Computing and Informatics, Vol. 32, 2013, 1013–1037

# AN EFFICIENT VISUAL ANALYSIS METHOD FOR CLUSTER TENDENCY EVALUATION, DATA PARTITIONING AND INTERNAL CLUSTER VALIDATION

Puniethaa Prabhu

*Department of Master of Computer Application*
*K. S. Rangasamy College of Technology, Tiruchengode*
*Namakkal (DT) – 637 215, Tamil Nadu, India*
*e-mail:* spunitha156@yahoo.co.in

Karuppusamy Duraiswamy

*Department of Computer Science and Engineering*
*K. S. Rangasamy College of Technology, Tiruchengode*
*Namakkal (DT) – 637 215, Tamil Nadu, India*
*e-mail:* drkduraiswamy@yahoo.co.in

**Abstract.** Visual methods have been extensively studied and performed in cluster data analysis. Given a pairwise dissimilarity matrix $D$ of a set of $n$ objects, visual methods such as Enhanced-Visual Assessment Tendency (E-VAT) algorithm generally represent $D$ as an $n \times n$ image $I(\overline{D})$ where the objects are reordered to expose the hidden cluster structure as dark blocks along the diagonal of the image. A major constraint of such methods is their lack of ability to highlight cluster structure when $D$ contains composite shaped datasets. This paper addresses this limitation by proposing an enhanced visual analysis method for cluster tendency assessment, where $D$ is mapped to $D'$ by graph based analysis and then reordered to $\overline{D'}$ using E-VAT resulting graph based Enhanced Visual Assessment Tendency (GE-VAT). An Enhanced Dark Block Extraction (E-DBE) for automatic determination of the number of clusters in I($\overline{D'}$) is then proposed as well as a visual data partitioning method for cluster formation from $I(\overline{D'})$ based on the disparity between diagonal and off-diagonal blocks using permuted indices of GE-VAT. Cluster validation measures are also performed to evaluate the cluster formation. Extensive

experimental results on several complex synthetic, UCI and large real-world data sets are analyzed to validate our algorithm.

**Keywords:** Visual clustering, graph analysis, cluster assessment tendency, automatic clustering, visual data partitioning and validation measures.

## 1 INTRODUCTION

A major issue in the data mining area is how to categorize the observed data into meaningful structures. Clustering analysis, also called segmenation analysis or taxonomy analysis, intends to identify homogeneous objects into a set of groups, named clusters, by given criteria. Partitioning the set of objects $O = \{o_1, o_2, \ldots, o_n\}$ into $C$ self-related objects is the major process of cluster analysis. Various clustering algorithms are reported in the literature [1, 2, 3]. All clustering algorithms will be analysed and subjective ($1 \leq C \leq n$) clusters numbers, even if no "definite" clusters exist, i.e., $C_1 \ldots C_c$, so that $C_i \cap C_j =$ if $i \neq j$ and $C_1 \cup C_2 \cup \ldots \cup C_c = O$. There have been large numbers of data clustering algorithms in the recent literature [1]. The general problems involved in clustering of unlabeled data sets are:

a) assessing cluster tendency, i.e., value of $C$,

b) grouping the data into $C$ meaningful sets, and

c) evaluating the Clusters discovered.

Given "only" a pairwise dissimilarity proximity matrix $D \in R^{n \times n}$ representing a data set of $n$ objects, this paper addresses these three problems, i.e., determining the number of clusters $C$ prior to clustering, partitioning the data into $C$ clusters and validating the clusters.

Most clustering algorithms desire the number of clusters $C$ as a key factor, so the quality of the resultant clusters mainly depends on the assessment of $C$. For several applications, users can choose the number of clusters with domain information. However, in various situations, the significance of $C$ is unknown and has to be predicted from the data itself. Diverse postclustering measures of cluster validity have been proposed to approximate $C$, e.g., [4, 5, 6, 7, 8, 9, 10], by choosing the best partition from a set of alternative partitions. To compare, cluster tendency assessment attempts to estimate $C$ before clustering occurs. Visualization used in cluster analysis maps the high-dimensional data to a 2D space and aids users having an intuitive and easy-to-understand graph/image to expose the grouping relationship among the data. In particular, the depiction of data structures in an image format has an extensive and continuous history, e.g., [11, 12, 13, 14, 15, 16]. The visual illustration of pairwise dissimilarity information about a set of $n$ objects is typically depicted as an $n \times n$ image, where the objects are reordered so that the resultant image is capable to highlight potential cluster structure in the data. A "useful"

reordered dissimilarity image (RDI) highlights possible clusters as a group of "dark blocks" along the diagonal of the picture, and can thus be viewed as a visual assist to tendency assessment.

Reordered dissimilarity images are generated by using any of the existing schemes in [12, 13, 14, 15, 16, 17]. For compactness, this paper focuses on one method for generating RDIs, namely Enhanced-Visual Assessment of cluster Tenedency (E-VAT) of Puniethaa [17], although the proposed approach can also be applied to any method that generates RDIs. However, several practical applications involve datasets with highly complex structure, which invalidate the assumption of compact, well-separated clusters. This paper proposes a new approach for generating RDIs that combines E-VAT with weighted graph analysis of pairwise data.The resulting Graph Enhanced-VAT (GE-VAT) image clearly demonstrates the number of clusters $C$ and the estimated sizes of each cluster for data sets with highly irregular cluster structures. Based on GE-VAT, the cluster structure in the data can be consistently estimated by visual assessment. An effective strategy to measure the "goodness" of GE-VAT images for automatically determining the number of clusters $C$ using Enhanced Dark Block Extraction (E-DBE) is also proposed. Also, a visual data partitioning algorithm based on the GE-VAT image and its unique block-structured property to cluster the data into $C$ groups is performed. By integrating cluster tendency assessment and cluster formation using RDI, the paper proposes a natural environment for visual cluster internal validation and interpretation. A wide range of primary and comparative experiments on synthetic, UCI and real-world data sets exhibit the efficacies of the proposed algorithms.

In summary, the major contributions of this paper comprise:

1. the GE-VAT algorithm for better revealing the hidden cluster structure of complex shaped data sets;
2. the efficient "goodness" measure of the GE-VAT images for automatic assessment of cluster tendency by E-DBE algorithm;
3. the valuable visual data partitioning algorithm is based on the GE-VAT images and
4. the visual cluster internal validation is performed on the cluster formation.

Schematic diagram of the overall proposed system and the major steps are shown in Figure 1. The rest of the paper is organized as follows: Section 2 illustrates the proposed graph based GE-VAT algorithm. Section 3 presents the strategy of E-DBE for automatically determining the number of clusters $C$ from the GE-VAT images, Section 4 shows how to find the $C$ clusters from the GE-VAT image using visual data partitioning algorithm. Section 5 extends the GE-VAT algorithm to evaluate the created cluster objects using internal cluster validation methods. Experimental results are shown in Section 6, prior to discussion and conclusion in Section 7.

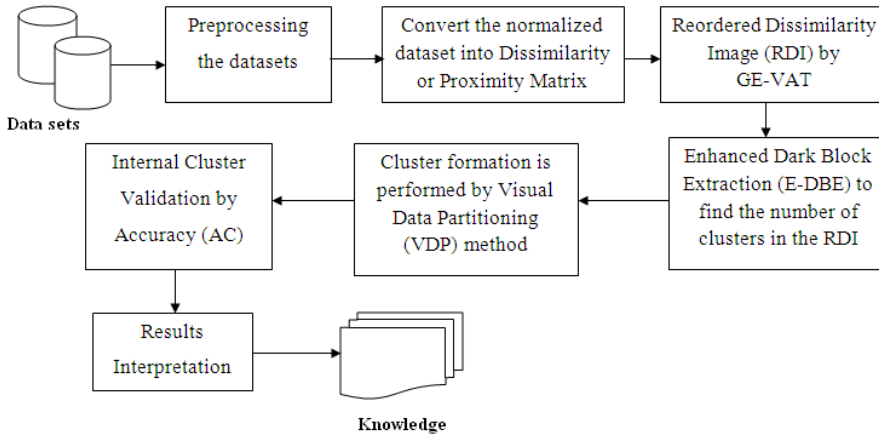



Figure 1. Automatic clustering procedure. These steps are related to each other and perform cluster partition.

## 2 GRAPH BASED ENHANCED VISUAL ASSESSMENT OF CLUSTER TENDENCY

The proposed work is built upon the E-VAT algorithm [17] (see Appendix). It simply reorders the data to reveal its hidden structure, which can be viewed as illustrative data visualization for estimating the number of clusters former to clustering. However, hierarchical structure could be detected from the reordered matrix if the diagonal sub-blocks are presented within larger diagonal blocks. The viewer can approximate the number of clusters $C$ from the E-VAT image by counting the number of dark blocks along the diagonal if these dark blocks acquire visual precision. Though, this is not always feasible. Dark block appears in the E-VAT image only when a fixed (or ellipsoidal) group exists in the data. For complex-shaped data sets where the borders between clusters become less distinct due to either significant overlay or rough geometries between different clusters, the resulting E-VAT images will degrade (see Figures 4 a), 5 a) and 6 a) for examples). Consequently, viewers may infer different numbers of clusters from such poor images, or even cannot estimate $C$ at all. This obviously raises a problem of whether to transform $D$ into a new form $R'$ so that the E-VAT image of $R'$ can become clearer and more helpful about the cluster structure. The proposed algorithm addresses this problem by combining the E-VAT algorithm with weighted graph analysis of the proximity matrix of the data.

Recently, a number of researchers have used graph analysis in applications such as random-walk [18], dimensionality reduction [19], image segmentation [20, 21] and data clustering [23]. These visual graph techniques commonly use the eigenvectors of a graph adjacency (or Laplacian matrix) to create a geometric representation of the

graph. Different methods are strongly connected, e.g., Laplacian eigenmaps [19] are much related to the mapping process used in spectral clustering algorithm depicted in [24]. Let $G(V, E, W)$ be a weighted undirected graph, where $V$ is a set of $n$ vertices (e.g., corresponding to $n$ objects $\{o_1, o_2, \ldots, o_n\}$ to be analyzed), $E = [e_{ij}]$ is the edge set with $e_{ij} = 1$ showing that there is an association between vertices $i$ and $j$ and 0 otherwise, and $W = [w_{ij}]$, an $n \times n$ affinity matrix or weighted adjacency matrix, includes the edge weights, with $w_{ij}$ representing the relation of the edge linking vertices $i$ and $j$. Most graph representation methods differ in terms of constructing the graph (reflected in $E$, e.g., the $\epsilon$ – neighborhood graph, the $K$-nearest neighbor graph [22] and the fully connected graph), weighting functions (reflected in $W$, e.g., simple 0-1 weighting and the normally used Gaussian similarity function) or graph Laplacians (e.g., the unnormalized Laplacian matrix $L = M - W$ and the normalized version $L = M^{-1/2}LM^{-1/2}$, where $M$ is a diagonal degree matrix of $G$, i.e., $m_{ij} = \sum_{j=1}^{n} w_{ij}$). The graphical decomposition of the Laplacian matrix provides useful information about the properties of the graph. It has been revealed experimentally that groups in the original data space may not correspond to curved regions, but once they are mapped to a spectral space spanned by the characteristic vectors of the Laplacian matrix, they are more likely to be changed into compact clusters [23, 25]. Based on this study, $D$ is embedded in a $k$-dimensional spectral space, where $k$ is the number of characteristics vectors used, such that each original data point is absolutely replaced with a new vector instance in this new space. After a complete study of recent spectral methods, combinations of graph Laplacian for obtaining a better graph embedding (and thus, better GE-VAT images, see Figures 4 b), 5 b) and 6 b) for example) are executed. GE-VAT algorithm and its pseudocode are summarized in Tables 1 and 2, respectively.

| |
|---|
| **Input:** $D = [d_{ij}]$: an $n \times n$ scaled matrix of pairwise dissimilarites $k$: the number of characteristic vectors used. |
| **Process**<br>**Step (1):** Compute a local scale parameter $\sigma_i$ for object $O_i$ using $\sigma_i = d(O_i, O_k) = d_{ik}$ where $O_k$ is the $K^{\text{th}}$ nearest neighbor of $O_i$. |
| **Step (2):** Build the weighting matrix $W \in R^{n \times n}$ by defining $W_{ij} = \exp(-d_{ij}d_{ji}/(\sigma_i\sigma_j))$ for $i \neq j$, and $w_{ii} = 0$. |
| **Step (3):** Construct the normalized Laplacian matrix $L' = M^{-1/2}WM^{-1/2}$. |
| **Step (4):** Choose the $K$ largest characteristic vectors of $L'$ to form the matrix $V = [v_1, \ldots, v_k] \in R^{n \times k}$ by stacking the characteristic vectors in columns. |
| **Step (5):** Normalize the rows of $V$ with unit Euclidean norm to generate $V'$. |
| **Step (6):** For $i = 1, 2, 3 \ldots n$ let $u_i \in R_k$ be the vector consequent to the $i^{\text{th}}$ row of $V'$ and treat it as a new instance (related to $O_i$). Then construct a new pairwise dissimilarity matrix $D'$ between instances. |
| **Step (7):** Apply the E-VAT algorithm to $D'$ to obtain the image $I(\overline{D'})$ |
| **Output:** Graph-mapped and reordered dissimilarity matrix $(\overline{D'})$ and its corresponding scaled gray-scale image $I(\overline{D'})$ |

Table 1. Graph based enhanced – visual assessment tendency algorithm

**Start**
**Input:** $D$ a scaled dissimilarity matrix, $K$ the number of characteristic vectors
**Output:** Reordered dissimilarity matrix ($\overline{D'}$) and gray scale image $I(\overline{D'})$
**Execute:** Graph based Enhanced – Visual Assessment Tendency Algorithm (GE-VAT)
While *objref* do
  For each profile *objref,k*
    $\sigma_{objref} \leftarrow$ distance$(O_{objref}, O_k)$ where $O_k$ is the $K^{\text{th}}$ nearest neighbor of $O_{objref}$
  End for
endwhile
// weighting Matrix $M$-Diagonal Degree Matrix and $W$-Adjency Matrix
For each profile $i$
  For each profile $j$
    If $i \neq j$ then $W_{ij} = \exp(-d_{ij}d_{ji})/\sigma_i\sigma_j$ else $W_{ii} = 0$
  endfor
endfor
For each profile $i$
  For each profile $j$
    $d_i \leftarrow W_{ij} + d_i$
  endfor
endfor
Construct the normalized Laplacian Matrix $L'$
  $L' \leftarrow M^{-1/2}WM^{-1/2}$
Compute $K$ largest characteristic vector for $L'$ to form the vector norm $V = [v_1, v_2, \ldots v_k]$ and normalize to $V'$.
For each profile $i$
  New dissimilarity matrix $D'$
endfor
Function EVAT$(D', \text{size}(D'), P_i)$ // $P_i$ – Permutation Vector
  Compute the reordered dissimilarity image $I(\overline{D'}) \leftarrow (\overline{D'})$
Return(RDI, $(\overline{D'}), I(\overline{D'})$)
End

Table 2. Pseudo code for graph based enhanced – visual assessment tendency algorithm

Some points about this algorithm are noted as follows:

- Using a definite local scaling parameter allows fine-tuning of the object-to-object distance according to the local statistics of the neighborhood surrounding objects $i$ and $j$, resulting in high likeness within clusters and low likeness across clusters, which has been demonstrated in [25], is a benefit for clustering.
- Performing the normalized Laplacian matrix $L' = M^{-1/2}WM^{-1/2}$.
- The computational complexity of the GE-VAT algorithm depends mainly on three parts, i.e., computing the local scale parameter $\sigma_i$, the characteristic decomposition of the normalized Laplacian matrix $L'$, and performing the E-VAT

algorithm. The corresponding runtime complexities for these three parts are, respectively, $O(Kn^2)$, $O(n^3)$ and $O(n^2)$. Thus, the total computational complexity of GE-VAT algorithm is $O(n^3 + (K+1)n^2)$.

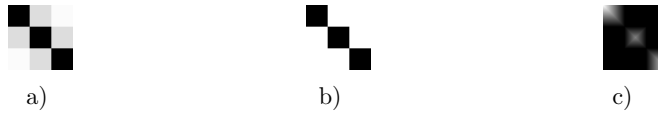
## 3 AUTOMATIC CLUSTER TENDENCY ASSESSMENT

Clustering in unlabeled data $O$ is the assignment of labels to the objects in $O$, where two essential components are needed; they are the number of groups to seek $C$ and a partitioning method to discover the $C$ clusters. In this section, the problem of cluster tendency assessment is explored by GE-VAT. Before designing an automatic method for estimating the number of clusters from the GE-VAT images, the characteristics of the GE-VAT images are examined. Figures 4 a), 5a and 6a show the examples for original E-VAT image and best (based on $k$ value) GE-VAT images with different numbers of characteristic vectors are shown in Figures 4 b), 5 b) and 6 b) respectively, GE-VAT images are generally clearer than the orginal E-VAT image in revealing real data structure.

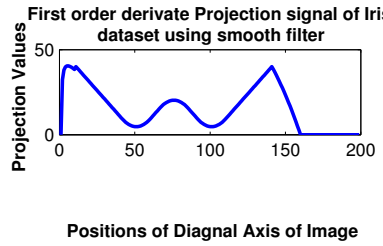
To enable automatic determination of the number of clusters, the "best" GE-VAT images in terms of "clarity" and "block structure" must be found. Each of the "block regions" in the image corresponds to either intracluster or intercluster dissimilarity values, while "clarity" is relevant to the degree of brightness difference of such blocks. For determining the number of clusters from the given GE-VAT image an Enhanced- Dark Block Extraction (E-DBE) is proposed based on [26] which relies on distance measure and basic image and signal processing techniques [27]. The procedure for E-DBE is summarized in Table 3. Figure 2 shows the results of E-DBE algorithm on iris dataset. In Figure 2 a) three dark blocks are presented on the diagonal which means there are three clusters in the data set for which the input is the outcome of GE-VAT algorithm. Next the binary GE-VAT image is revealed in Figure 2 b). Distance transformation on the binary image is performed to obtain the gray-scale image which is shown in Figure 2 c). Later the positions of diagonal values in the gray-scale image are projected as a first order derivative of iris data set using smooth, moving, sgolay filters are shown as a projection signal in the Figures 2 d), 2 e) and 2 f), respectively. The peaks in the projection signals the number of clusters presented in the data set (here $C = 3$).

Some points about this algorithm are noted as follows:

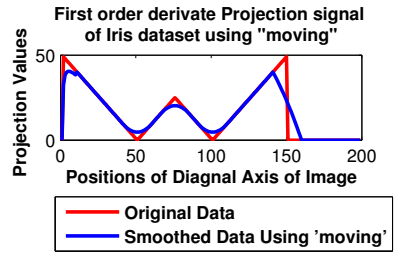
- Transform the original matrix $D$ to a new dissimilarity matrix $D'$ using a "monotonic" exponential function $f(v) = 1 - \exp(-v/\sigma)$ (parameter $\sigma$ may be merely selected as the global threshold significance obtained by Otsu's algorithm [28]).
- Adaptive threshold algorithm to obtain a new threshold $\sigma'$ to convert the binary image by $\text{Image}_{ij}^{(2)} = 1$ if $\text{Image}_{ij}^{(1)} > \sigma'$ and $\text{Image}_{ij}^{(2)} = 0$ otherwise.
- Distance transformation is a form of depiction of a digital image, which converts a binary image to a gray-scale image in which the value of each pixel is the distance from the pixel to the adjacent non-zero pixel in the binary $\text{Image}^{(2)}$.



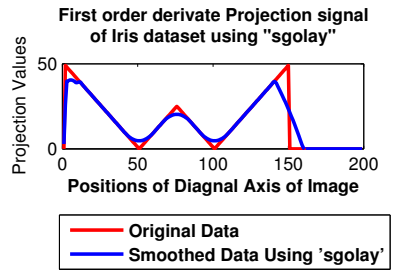
a) b) c)




d)




e)




f)

Figure 2. Sample results of the E-DBE algorithm on iris dataset; a) GE-VAT image of iris dataset; b) binary GE-VAT image of iris dataset; c) distance transformed image; d) first order derivative projection signal obtained using "smooth"; e) first order derivative projection signal obtained using "moving"; f) first order derivative projection signal obtained using "sgolay"

| |
|---|
| **Input:** $n \times n$ – scaled matrix of dissimilarities $D = [d_{ij}]$, the proportion of the allowed minimum cluster size of the data size $n$. |
| **Process**<br>**Step (1):** Transform $D$ to a new dissimilarity matrix $D' = d_{ij} = 1 - \exp(-d_{ij}/\sigma)$.<br>$\sigma$ – scale parameter determined $D$ using Otsu [28] automatically. |
| **Step (2):** Form a RDI Image$^{(1)}$ corresponding to $D'$ using GE-VAT algorithm. |
| **Step (3):** Threshold the Image$^{(1)}$ to obtain binary Image$^{(2)}$ using the adaptive threshold [29] algorithm. |
| **Step (4):** Perform a distance transform on Image$^{(2)}$ to obtain a new gray-scale Image$^{(3)}$, and scale the pixel values to $[0, 1]$. |
| **Step (5):** Project the pixel values of the Image$^{(3)}$ onto the main diagonal axis of the image to form a projection signal Histogram$^{(1)}$. |
| **Step (6):** Filtering the projected signal is performed by Savitzky-Golay filter design [30]. |
| **Step (7):** Compute the first order derivative of the Histogram$^{(1)}$ to obtain signal Histogram$^{(2)}$. |
| **Output:** The number of dark blocks $C$ (i.e., count the number of major peaks)<br>presented in the RDI. |

Table 3. Enhanced dark block extraction algorithm

- Savitzky-Golay filters are optimal, they minimize the least-squares error in fitting a polynomial to frames of noisy data and find the original sets of peaks $p_i$ and valleys $v_j$ by finding the equivalent from-positive-to-negative zero-crossing points and from-negative-to-positive zero-crossing points.
- The computational complexity of this algorithm mainly depends on the computation of gray scale images and optimal thresholds, which is $O(n^2)$.

## 4 VISUAL DATA PARTITIONING

In this section, GE-VAT is further explored for the problem of visual data partitioning. That is, whether the method can automatically extort a crisp $C$-partition of $O$ directly from the visual facts in Image($\overline{D'}$)? If so, how well does it execute? In common, the $C$-partitions of a data set $O$ are sets of $c.n$ values $u_{ik}$ that can be easily arrayed as $c \times n$ matrix $U = [u_{ik}]$. The set of all nondegenerate $c$-partition matrices for $O$ is

$$\left[H_{hcm} = \left\{U \in R^{c \times n} | 0 \leq u_{ik} \leq 1, \forall i, k\right\}\right] \tag{1}$$

with

$$\sum_{i=1}^{c} u_{i,k} = 1, \forall k \quad \text{and} \quad \sum_{k=1}^{n} u_{i,k} > 0, \forall 1. \tag{2}$$

Element $u_{ik}$ of $U$ is the membership of object $k$ in clusters $i$. In the case of a "crisp" (or hard) partition (not fuzzy), $u_{ik} = 1$ if $O_k$ is labeled 1 and 0 otherwise. The significant property of Image($\overline{D'}$) is that it has, starting in the upper left corner, dark blocks along its main diagonal. Consequently, constrain the search through $H_{hcn}$ to those partitions that imitate the block structure in Image($\overline{D'}$) [31], i.e.,

$$H^*_{hcm} = \{U \in H_{hcm}\}. \tag{3}$$

Let $U$ in $H^*_{hcn}$ be an aligned $C$-partition of $O$ when its entries form $C$ contiguous blocks of 1 s in $U$, ordered to initiate from the upper left corner, and continue down and to the right. Every member of $H^*_{hcn}$ is isomorphic to the distinctive set of $C$ discrete integers, i.e., the cardinalities of the $C$ clusters in $U$ that gratify $\{\{n_1 | 1 \le n_1; 1 \le i \le C; \sum_{i=1}^{c} n_i = n\}\}$, so associated partitions can be alternatively specified by $\{n_1 : n_2 : \ldots : n_c\}$. The important uniqueness of Image($\overline{D'}$) that can be exploited for finding a high-quality candidate partition $U$ are the contrast differences between the dark blocks along the main diagonal and the pixels closest to them.

| |
|---|
| **Input:** Image($\overline{D'}$): the GE-VAT image generated from a set of n objects.<br>$\pi()$**:** the permutation index obtained during E-VAT re-ordering<br>$C$: the number of clusters |
| **Process**<br>**Step (1):** Perform a global threshold on Image($\overline{D'}$)$^{(1)}$ to obtain a binary image Image($\overline{D'}$)$^{(2)}$. |
| **Step (2):** Apply Distance Transformation (DT) on binary image Image($\overline{D'}$)$^{(2)}$. For each pixel in $\overline{D'}$, the distance transformed $\overline{D''}$ assigns a number that is the distance between that pixel $\overline{D'}$ and the nearest nonzero pixel of $'$. Here the "city block" distance measure is used to execute the DT. |
| **Step (3):** Obtain the diagonal values of $\overline{D''}$ and convert into a $1D$ vector norm $V*$. |
| **Step (4):** Transform the non-zero elements of $V*$ into cluster partition $U*$ (which is equivalent to obtaining the sizes of each cluster $\{n_1, \ldots, n_c\}$). The position $p_1$ of the first '1' in $V*$ means the first cluster partition is from sample 1 to $p_1$. The position $p_j$ $(j = 2, \ldots, C-1)$ of the $j^{\text{th}}$ '1' means the $j^{\text{th}}$ cluster partition is from sample $(p_{j-1}+1)$ to $p_j$. The $C^{\text{th}}$ cluster partition is from sample $(p_{c-1}+1)$ to $n$. |
| **Step (5):** Retrieve real object indices in each cluser $C_i$ with the permutation index $\pi()$, i.e., $C_1 = \{O_{\pi(1)}, \ldots, O_{\pi(n_1)}\}$ and $C_i = \{O_{\pi(n_{i-1}+1)}, \ldots, O_{\pi(n_{i-1}+n_i)}\}$ for $i = 2, \ldots, C$. |
| **Output:** The data partitioning results $\{C_1, \ldots, C_c\}$. |

Table 4. Visual data partitioning (VDP) algorithm

The proposed algorithm aims to generate candidate partitions in $H^*_{hcn}$ by testing their robustness to the clusters suggested by the aligned dark blocks in Image($\overline{D'}$) (i.e., dissimilarites in non-dark blocks off-diagonal) [31]. The proposed visual data

partitioning procedure is based on the GE-VAT image which is summarized in Table 4.

Several points about this algorithm are noted as follows:

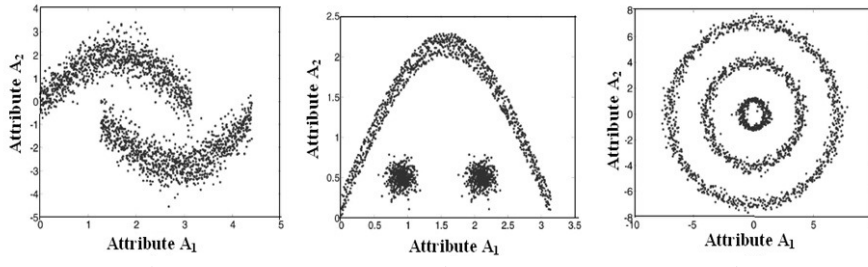
- Alternatively the proposed algorithm can be executed using optimization algorithms such as genetic algorithm and particle swarm optimization.
- In addition to the GE-VAT image, this visual partitioning procedure can also operate on other reordered dissimilarity images.
- Sampling-based extended scheme can be applied for large object data sets.
- The time complexity of this algorithm mainly depends on the distance tranformation, object size, i.e., space complexity $O(n^2)$ and vector norm $V*$.

## 5 EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithms, a number of experiments have been made on synthetic, UCI and real world datasets (summarized in Table 5). The preprocessing data mining techniques such as mean imputation method [32] for missing data and Z-score normalization [33] are performed on data sets to compute the distance matrix $D$, where $D$ is the input dissimilarity matrix for GE-VAT algorithm. All experiments were implemented in a Matlab 7.1.0.246 environment on a PC with an Intel 2.10 GHz CPU and 3 GB memory running Windows XP.

### 5.1 Review of Data Sets

Three synthetic data sets with diverse data structure (i.e., dissimilar numbers of clusters and different data distributions) are used in the experiments. The scatter plots of these synthetic data sets are shown in Figure 3, in which each one corresponds to a visually meaningful collection. These three data sets are taken from [36]; the data sets involve more irregular and dense data structures, in which an observable cluster centroid for each group is not predictable. These data sets include different scales between clusters or some clusters are concealed in a cluttered background.




a) b) c)

Figure 3. Scatter plot of synthetic data sets (S-1, S-2 and S-3)

Six real-world data sets were also considered to assess the proposed algorithms, five of which are from the UCI machine Learning Repository and one from various Integrated Counseling and Testing Center (ICTC) and Antiretroviral (ART) centers of Tamilnadu and Pondicherry. i.e., U-1, U-2,. . . , U-5and R-1. In concise, U-1 (Dermatology) database includes 357 instances, each of which has 33 attributes and belongs to one of 6 classes. U-2 (Heart) data set from [34] includes 270 instances, each of which has 12 attributes and belongs to one of 2 classes. U-3 (Hepatisis) data set is a $72 \times 72$ matrix consisting of pairwise dissimilarities from a set of 72 patients clinical report that were clustered into 2 groups (alive or dead) which had 19 immunological parameters. U-4 (iris) data set contains 3 physical classes, 50 instances each, where each class refers to a type of iris plant with 4 attributes. U-5 (wine) data set includes 178 samples from 3 wine cultivators with 12 various attributes.

R-1 (HIV/AIDS) diagnosis data set contains 400 objects. The attributes are Age, Sex, WT, HB, Treat Drug, Pill count, Initial drug, Occupation, Marital status, CD4, CD8, Ratio, WBC, RBC, PCV, platelet, TLC, SGPT, SGOP and Drug regimen- Class Attribute (CA). The entire integer of items in this data set is $n = 400$.

| **Data Set** | **Physical Class** ($C_p$) | **# attributes** | **Size** ($n$) | Number of Clusters $C$ | | |
|---|---|---|---|---|---|---|
| | | | | **Manual E-VAT** ($C^m_{EV}$) | **Manual GE-VAT** ($C^m_{GEV}$) | **Automatic E-DBE** ($C^a_{EDBE}$) |
| S-1 | 2 | 2 | 2 000 | $\geq 1$ | 2 | 2 |
| S-2 | 3 | 2 | 2 000 | $\geq 2$ | 3 | 3 |
| S-3 | 3 | 2 | 2 500 | $\geq 1$ | 3 | 3 |
| U-1 | 6 | 33 | 357 | 2 | 6 | 6 |
| U-2 | 2 | 12 | 270 | $\geq 1$ | 2 | 2 |
| U-3 | 2 | 19 | 72 | $\geq 2$ | 2 | 2 |
| U-4 | 3 | 4 | 150 | $\geq 2$ | 2 | 2 |
| U-5 | 3 | 12 | 178 | $\geq 1$ | 3 | 3 |
| R-1 | 6 | 19 | 400 | $\geq 1$ | 6 | 6 |

Table 5. Summary of data sets used and estimating $C$ results

### 5.2 Determining the Number of Clusters

For each data set data mining preprocessing techniques, pairwise dissimilarity matrix $D$ are computed in the original attribute space. The E-VAT images are shown in Figure 4 a) for the synthetic data sets, Figure 5 a) for UCI data sets and Figure 6 a) for real data set. It is viewed that the cluster structure of the data in these E-VAT images is not clearly highlighted. Consequently, viewers have complexity in giving a sound result about the number of clusters in these data sets; different viewers may infer diverse estimates of $C$. Later, GE-VAT algorithm is applied to each of the data

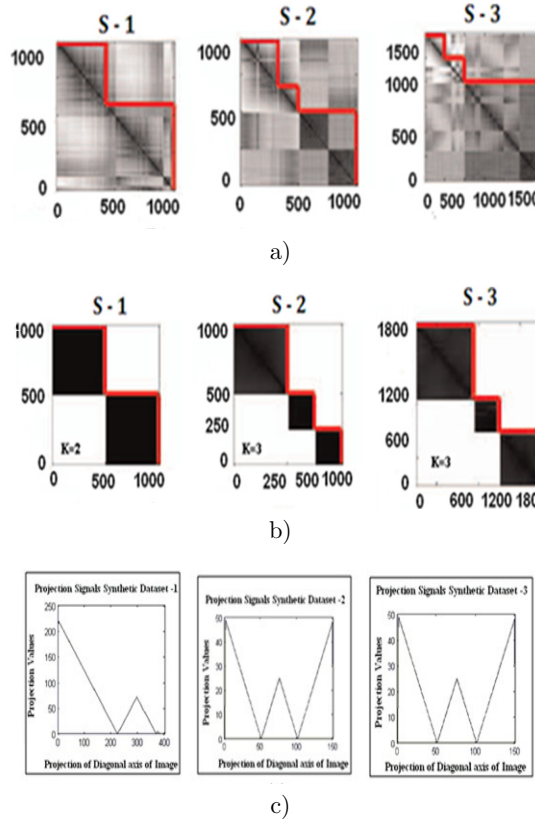
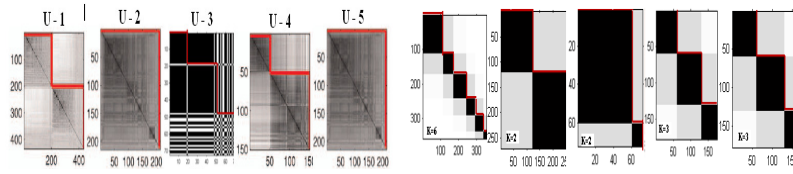


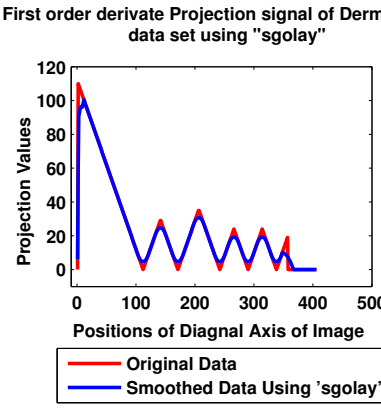

Figure 4. a) original E-VAT images of three synthetic data sets with visual clustering results shown by dark red lines; b) the corresponding best GE-VAT images with visual clustering results; c) the corresponding E-DBE projection signals to determine the number of clusters automatically

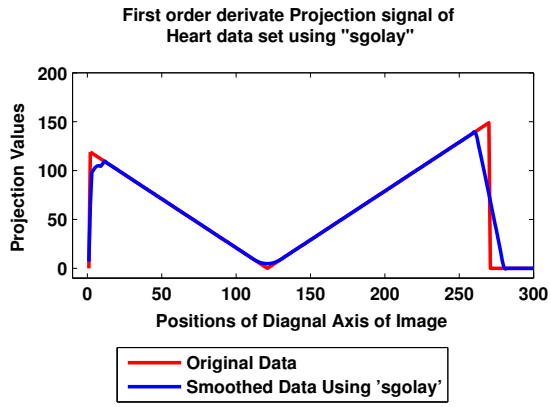
sets and the results for synthetic, UCI and real data sets are shown in Figures 4 b), 5 b) and 6 b), respectively.

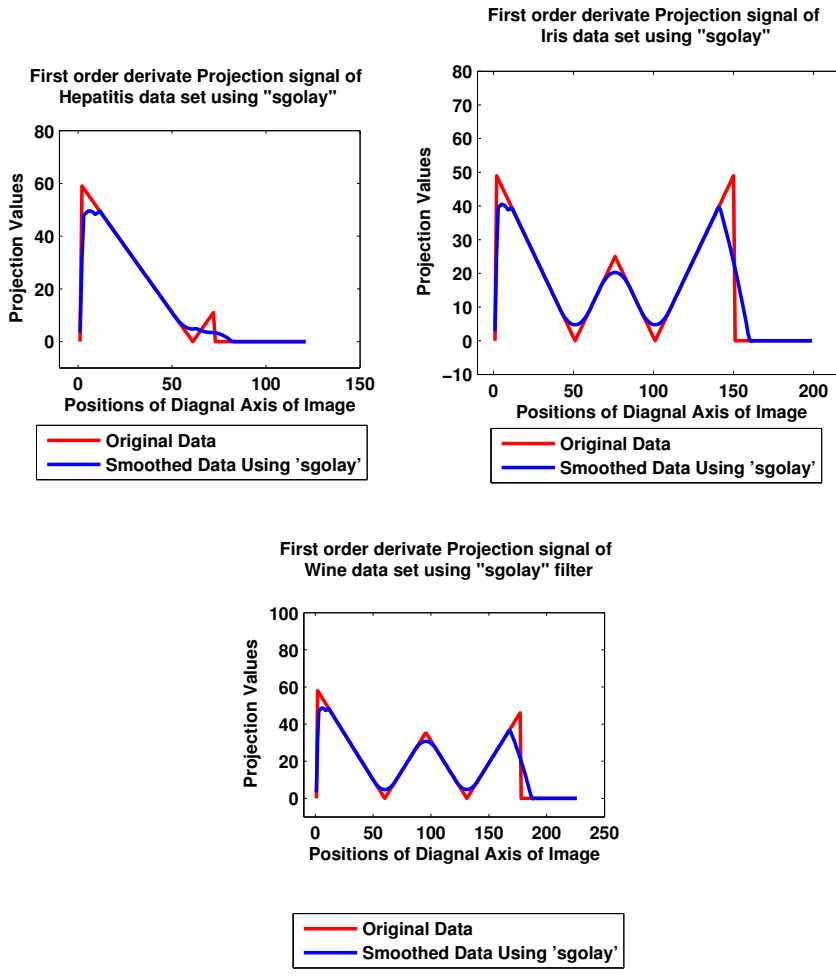
In contrast to the original E-VAT images (Figures 4 a), 5 a) and 6 a)), the GE-VAT images (Figures 4 b), 5 b) and 6 b)) generally have clearer presentations of the block structure on the diagonal and thus better highlight the concealed cluster structure. Table 5 summarizes the number of clusters determined by manual inspection from the original E-VAT image ($C^m_{EV}$), manual inspection from a series of GE-VAT images ($C^m_{EV}$) and automatic determination of cluster numbers from GE-VAT images using E-DBE ($C^m_{EDBE}$). The results of cluster number assessment from the GE-VAT images for the 9 data sets are correct in terms of the number of actual physical classes $C_p$; this was estimated by manual inspection of the data sets. The


U-1
U-2
U-3
U-4
U-5

a) b)

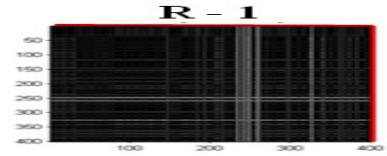

First order derivate Projection signal of Dermatology data set using "sgolay"
Projection Values
Positions of Diagnal Axis of Image
Original Data
Smoothed Data Using 'sgolay'


First order derivate Projection signal of Heart data set using "sgolay"
Projection Values
Positions of Diagnal Axis of Image
Original Data
Smoothed Data Using 'sgolay'

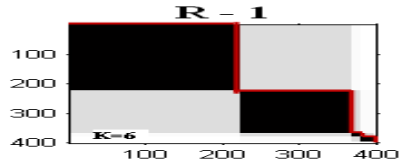



c)

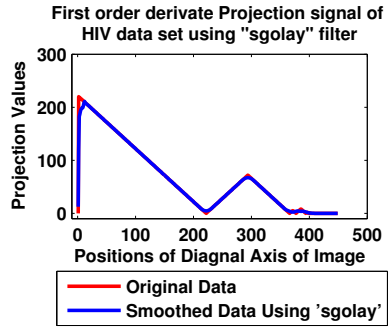
Figure 5. a) original E-VAT images of five UCI data sets with visual clustering results shown by dark red lines; b) the corresponding best GE-VAT images with visual clustering results; c) the corresponding E-DBE projection signals to determine the number of Clusters




a)




b)




c)

Figure 6. a) original E-VAT images of real HIV data set with visual clustering results shown by dark red lines; b) the corresponding best GE-VAT images with visual clustering results; c) the corresponding E-DBE projection signals to determine the number of clusters automatically

outcome again highlights the benefits of converting $D$ to $D'$ by graph embedding to obtain a more accurate estimate of $C$. Table 6 shows the concise account of cluster results of VDP algorithm for data sets using GE-VAT images.

### 5.3 Efficient Visual Data Partitioning (vdp) and Algorithm Comparison

Cluster validations, which assess the goodness of clustering outcomes [5], have long been recognized as one of the vital issues to the success of clustering applications. External and internal clustering validations are the two main categories of cluster validation. External validation measure is entropy, which evaluates the "purity" of

| Data Set | # of clusters through VDP | Cluster | | Manual inspection of class |
|---|---|---|---|---|
| | | Cluster # | # of objects in each cluster | |
| S-1 | 2 | 1 | 1 000 | 2 |
| | | 2 | 1 000 | |
| S-2 | 3 | 1 | 1 000 | 3 |
| | | 2 | 500 | |
| | | 3 | 500 | |
| S-3 | 3 | 1 | 1 000 | 3 |
| | | 2 | 750 | |
| | | 3 | 250 | |
| U-1 | 6 | 1 | 110 | 6 |
| | | 2 | 59 | |
| | | 3 | 70 | |
| | | 4 | 48 | |
| | | 5 | 51 | |
| | | 6 | 19 | |
| U-2 | 2 | 1 | 150 | 2 |
| | | 2 | 120 | |
| U-3 | 2 | 1 | 60 | 2 |
| | | 2 | 12 | |
| U-4 | 3 | 1 | 50 | 3 |
| | | 2 | 50 | |
| | | 3 | 50 | |
| U-5 | 3 | 1 | 59 | 3 |
| | | 2 | 71 | |
| | | 3 | 48 | |
| R-1 | 6 | 1 | 221 | 6 |
| | | 2 | 144 | |
| | | 3 | 11 | |
| | | 4 | 17 | |
| | | 5 | 5 | |
| | | 6 | 1 | |

Table 6. Cluster results of Visual Data Partitioning (VDP) algorithm for data sets using GE-VAT images

clusters and internal validation measures only rely on information in the data [35]. In this paper the proposed visual partitioning algorithm evaluates by comparing the proposed algorithm with the ground truth label (available for these 9 data sets). Accuracy metric (AC) has been widely used for cluster validation [24, 37, 38]. Suppose that $M_i^C$ is the clustering label of an object $O_i$ and $M_i^G$ is the related ground truth label; then AC is defined as $Max_{map} \sum_{i=1}^{n} \delta(M_i^g, map(Z_i^C))/n$ where $n$ is the total number of objects in the data, $\delta(M_1, M_2)$ is the delta function that equals 1 if and only if $Z_1 = Z_2$ and 0 otherwise, and $map$ is the mapping function that permutes clustering labels to equivalent labels specified by the ground truth. The

| Data Set | Cluster # ($C$) | K-mean ($K_m$) | Ward ($L_w$) | GE-VAT$_\sigma$ | GE-VAT$_{\sigma i}$ | VDP$_{GE-VAT}$ | VDP$_{E-VAT}$ |
|---|---|---|---|---|---|---|---|
| S-1 | 2 | 88.30 | 71.70 | 100.0 | 100.0 | 100.0 | 90.15 |
| S-2 | 3 | 76.05 | 77.95 | 100.0 | 100.0 | 100.0 | 57.50 |
| S-3 | 3 | 45.64 | 48.83 | 100.0 | 100.0 | 100.0 | 53.21 |
| U-1 | 6 | 96.05 | 96.63 | 96.78 | 96.78 | 94.88 | 65.15 |
| U-2 | 2 | 88.14 | 91.72 | 87.13 | 88.05 | 90.80 | 83.45 |
| U-3 | 2 | 90.15 | 95.56 | 100.0 | 100.0 | 100.0 | 97.33 |
| U-4 | 3 | 81.45 | 89.33 | 90.67 | 93.33 | 98.67 | 67.33 |
| U-5 | 3 | 96.30 | 92.70 | 97.75 | 97.75 | 98.31 | 49.33 |
| R-1 | 6 | 76.19 | 95.55 | 82.15 | 81.60 | 96.00 | 68.95 |

Table 7. Comparison of clustering algorithm performance accuracy (percent)

| S. No | Proposed Methods | Complexity | Capability of tackling high dimensional data |
|---|---|---|---|
| 1 | GE-VAT | $O(n^3 + (K+1)n^2)$ | Yes |
| 2 | E-DBE | $O(n^2)$ | Yes |
| 3 | VDP | $O(n^2)$ | Yes |

Table 8. Computational complexity for the proposed algorithms

Kuhn-Munkres algorithm is usually used to obtain the best mapping [39]. Visualization of the proposed visual partitioning algorithm on synthetis, UCI and real data sets are shown in Figures 4 b), 5 b) and 6 b), respectively. The clustering validation of the proposed visual data partitioning algorithm on the orginal E-VAT image ($VDP_{E-VAT}$) and the GE-VAT ($VDP_{GE-VAT}$) are summarized in Table 7, from which the proposed algorithm obtains satisfactory partitioning results, i.e., $VDP_{GE-VAT}$ performs better than $VDP_{E-VAT}$.

Several typical clustering algorithms are also implemented for comparison. These algorithms are $K$-means ($K_m$), Ward's hierarchical clustering ($L_w$) [40], GE-VAT with global-scale parameter [24] (GE-VAT$_\sigma$) and GE-VAT with local-scale parameter [36] (GE-VAT$_{\sigma i}$). The clustering accuracies of these algorithms on these 9 synthetic, UCI and real data sets are listed in Table 7, from which it can be seen that overall precision of the proposed cluster partitioning algorithm on the GE-VAT image is better than that of $K$-means, Ward's algorithm and standard graph based clustering with local scale parameter is comparable to that of graph clustering with global scaling. In addition, visual methods give intuitive interpretation on the number of clusters, cluster structure and partition outcomes from the images, as well as eliminating the randomly initialized $K$-means clustering stage.

## 6 DISCUSSION AND CONCLUSION

This paper has offered an enhanced visual approach toward automatically determining the number of clusters and partitioning data in either object or pairwise rela-

tional form. In order to better reveal the hidden cluster structure, particularly for complex-shaped data sets, the E-VAT algorithm has been enhanced by using graph based analysis of the proximity matrix of the data. Based on GE-VAT the enhanced dark block extraction for automatically determining the number of clusters has been proposed. Later, visual clustering algorithm based on GE-VAT images have derived based on its block-structured property. A sequence of primary and comparative experiments on synthetic, UCI and real-world data sets have confirmed that our algorithms execute well in requisites of both visual cluster tendency assessment and data partitioning. The potentials of the proposed algorithm are applied in image segmentation, grouping of complex real-world datasets and in feature extraction.

There are strong relations between the GE-VAT algorithm and other works: both the GE-VAT algorithm and the spectral clustering algorithm described [24] use graph based normalized Laplacian matrix that is essentially for graph embeding procedure of [19]. A major property of the graph embedding framework is the entire preservation of the cluster structure in the embedding space. For novel representations in the embedding space, spectral clustering in [24] performs K-means to cluster them; while the proposed visual data clustering algorithm first converts them to an enhanced reordered dissimilarity (GE-VAT) image and then uses the diagonal values of the tranformed image by E-DBE with permutation index to partition its block structures. A local scaling method is suggested in [36] to change the global scale $\sigma$ in [24], leading to better clustering, mainly when the data includes manifold scales or when the clusters are positioned within cluttered background.

The proposed algorithms will possibly reach their useful limit when the image formed by any reordering of $D$ is not from a well-coherent dissimilarity matrix. The present method compares positively to postclustering validation methods in computational effectiveness. Next, the method does not eliminate the need for cluster validity (i.e., the third problem in cluster analysis). Computational complexity for the proposed algorithms is revealed in Table 8 which shows E-DBE and VDP has less computation time compared to GE-VAT. Cluster performance evaluation depends significantly on the choice of the validity criteria and the clustering algorithms. The existing index-based validation methods are apparently an approximate outcome to validate the proposed visual algorithms. The method for finding direct visual validation process will be one of the important issues in future work.

## A APPENDIX

### The Enhanced – VAT Algorithm

The E-VAT algorithm [17] works on a pairwise dissimilarity matrix.

Let $O = \{o_1, o_2, \ldots, o_n\}$ denote $n$ objects in the data and $D$ a pairwise matrix of dissimilarities between objects, each element of which, $d_{ij} = d(o_i, o_j)$, is the dissimilarity between objects $o_i$ and $o_j$ and usually satisfies $1 \geq_{ij} \geq 0$; $d_{ij} = d_{ji}$; $d_{ii} = 0$, for $1 \leq i, j \leq n$.

**Input**

Load the multi dimensional dataset and convert it into dissimilarity matrix using Euclidean, Hamming and Mahalanobis distance for numerical, categorical and mixed attributes respectively. Consider the dataset as $n \times n$ dissimilarity matrix.

**Process**

**Step (1):** Transform $D$ to a new dissimilarity matrix $R$ with $d'_{ij} = 1 - \exp(-d_{ij}/\sigma)$, where $\sigma$ is a scale parameter determined from D using the algorithm of Otsu [28] automatically.

**Step (2):** Form a RDI image $I^{(1)}$ corresponding to $R$ using the VAT algorithm. Let $I$ and $J$ be subsets of $K = \{1, \ldots, n\}$. We let argmin $\{R_{pq}\} p \in I$, $q \in J$ denote the set of *all* ordered index pairs $(i, j)$ in $I \times J$ such that $R_{ij} = \arg\min\{R_{pq}\} p \in I$, $q \in J$.

**Step (2.1):**

Let $I = \Phi$, $J = \{1, 2 \ldots n\}$ and $P = (0, \ldots 0)$.
Choose $(i, j) \in \arg_{p \in j \text{ and } q \in j} \max\{d_{pq}\}$
Place $P(1) = i$, $I \leftarrow \{i\}$ and $J \leftarrow J - \{i\}$

**Step (2.2):**

Iterate for $t = 2 \ldots n$
Select $(i, j) \in \arg_{p \in i \text{ and } q \in j} \min\{d_{pq}\}$
Set $P(t) = j$, revise $I \leftarrow I \cup \{j\}$ and $J \leftarrow J - \{j\}$

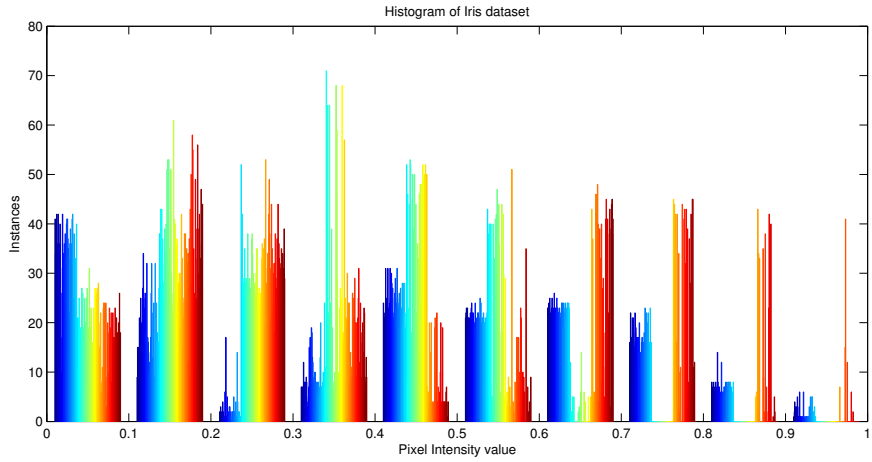
**Step (2.3):** Figure the dissimilarity template or matrix $R = [d_{ij}] = [d_{P(i)P(j)}]$
Where $1 \leq i, j \leq n$

**Step (3):** Display the reordered matrix $\tilde{R}$ as the ODI $\tilde{I}$ using the conventions given above.

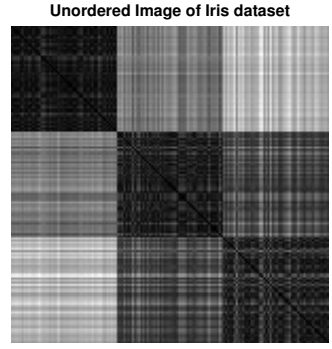
**Output**

Gray scale image $I(D)$, which denotes maximum $(d_{ij})$ to white and minimum $(d_{ij})$
to black.

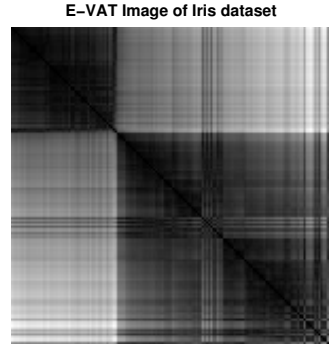
Table 9. The enhanced VAT algorithm




a)




b)




c)

Figure 7. Sample E-VAT algorithm for iris dataset a) Histogram of iris data set; b) Unordered image; c) Reordered E-VAT image I ($\overline{D'}$)

The E-VAT algorithm displays a reordered dissimilarity matrix of $D$ as a gray-scale image. The E-VAT algorithm is summarized in Table 9. The reordering idea is to find $P$ so that $\tilde{R}$ is as close to a block diagonal form as possible.

An example of E-VAT is shown in Figure 7. Histogram of iris data set of $n = 150$ points in $R^2$ is displayed in Figure 7 a), data points were converted to a $150 \times 150$ dissimilarity matrix $D$ by computing the distance measures based on the attribute characteristic between each pair of points. Figure 7 b) shows the unordered image of iris data set. In E-VAT image [13] in Figure 7 c) the dark blocks are not clearly visible, further reordering is necessary to reveal the underlying cluster structure of the data.

**Acknowledgments**

The authors express their sincere thanks to the Chairman and Principal of K. S. Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India for their support and constant encouragement and the anonymous reviewers for their valuable comments and kind suggestions for further improvement of the manuscript.

## REFERENCES

[1] Xu, R.—Wunsch, D. II.: Survey of Clustering Algorithms. IEEE Trans. Neural Networks, Vol. 16, 2005, No. 3, pp. 645—678.

[2] Jain, A. K.—Dubes, R. C.: Algorithms for Clustering Data. Prentice Hall 1998.

[3] Bezdek, J. C.—Keller, J. M.—Krishnapuram, R.—Pal, N. R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers 1999.

[4] Maulik, U.—Bandyopadhyay, S.: Performance Evaluation of Some Clustering Algorithms and Validity Indices. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, 2002, No. 12, pp. 1650–1654.

[5] Maimon, O.—Rokach, L.: Decomposition Methodology for Knowledge Discovery and Data Mining. World Scientific 2005, pp. 90–94.

[6] Bezdek, J. C.—Pal, N. R.: Some New Indices of Cluster Validity. IEEE Trans. System, Man and Cybernetics, Vol. 28, 1998, No. 3, pp. 301–315.

[7] Tibshirani, R.—Walther, G.—Hastie, T.: Estimating the Number of Clusters in a Data Set via the Gap Statistics. J. Royal Statistical Soc. B, Vol. 63, 2001, pp. 411–423.

[8] Calinski, R. B.—Harabasz, J.: A Dendrite Method for Cluster Analysis. Comm. In Statistics, Vol. 3, 1974, pp. 1–27.

[9] Dunn, J. C.: Indices of Partition Fuzziness and the Detection of Clusters in Large Sets. Fuzzy Automata and Decision Processes, Elsevier 1976.

[10] Hu, X.—Xu, L.: A Comparative Study of Several Cluster Number Selection Criteria. Intelligent Data Engineering and Automated Learning, Springer 2003, pp. 195–202.

[11] Cardona, M.—Colomer, M.-A.—Zaragoza, A.—Perez-Jimenez, M. J.: Hierarchical Clustering with Membrane Computing. Computing and Informatics, Vol. 27, 2008, No. 3+, pp. 497–513.

[12] Ling, R.: A Computer Generated Aid for Cluster Analysis. Comm. ACM, Vol. 16, 1973, pp. 355–361.

[13] Bezdek, J. C.—Hathaway, R. J.: VAT: A Tool for Visual Assessment of (Cluster) Tendency. Proc. Int'l Joint Conf. Neural Networks 2002, pp. 2225–2230.

[14] Tran-Luu, T.: Mathematical Concepts and Novel Heuristic Methods for Data Clustering and Visualization. Ph. D. thesis, Univ. of Maryland 1996.

[15] Rousseeuw, P. J.: A Graphical Aid to the Interpretations and Validation of Cluster Analysis. J. Computational and Applied Math., Vol. 20, 1987, pp. 53–65.

[16] Dhillon, I.—Modha, D.—Spangler, W.: Visualizing Class Structure of Multidimensional Data. Proc. $30^{\text{th}}$ Symp. Interface: Computing Science and Statistics 1998.

[17] Prabhu, P.—Duraiswamy, K.: Enhanced VAT for Cluster Quality Assessment in Unlabeled Datasets. Journal of Circuits, Systems and Computers, Vol. 21, 2012, No. 1, pp. 1–19.

[18] De Sterck, H.—Henson, V. E.—Sanders, G.: Multilevel Aggregation Methods for Small-World Graphs with Application to Random-Walk Ranking. Computing and Informatics, Vol. 30, 2011, No. 2, pp. 225–246.

[19] Belkin, M.—Niyogi, P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. Advances in Neural Information Processing Systems, MIT Press 2002.

[20] Weiss, Y.: Segmentation Using Eigenvectors: A Unifying View. Proc. IEEE Int'l Conf. Computer Vision 1999, pp. 975–982.

[21] Shi, J.—Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, 2000, No. 8, pp. 888–905.

[22] Khedr, A. M.: Nearest Neighbor Clustering over Partitioned Data. Computing and Informatics, Vol. 30, 2011, No. 5, pp. 1011–1036.

[23] von Luxburg, U.: A Tutorial on Spectral Clustering. Technical report, Max Planck Inst. for Biological Cybernetics 2006.

[24] Ng, A.—Jordan, M.—Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm. Advances in Neural Information Processing Systems. MIT Press 2002.

[25] Chung, F.: Spectral Graph Theory Vol. 92. Am. Math. Soc. 1997.

[26] Prabnu, P.—Duraiswamy, K.: Enhanced Dark Block Extraction Method Performed Automatically to Determine the Number of Clusters in Unlabeled Data Sets. International Journal of Computers, Communications and Control, Vol. 8, 2013, No. 2, pp. 275–293.

[27] Gonzalez, R. C.—Woods, R. E.: Digital Image Processing. Prentice Hall 2002.

[28] Otsu, N.: A Threshold Selection Method from Gray-level Histograms. IEEE Trans. Systems, Man, and Cybernetics, Vol. 9, 1979, No. 1, pp. 62–66.

[29] Sezgin, M.—Sankur, B.: Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. Journal of Electronic Imaging, Vol. 13, 2004, No. 1.

[30] Savitzky, A.—Golay, M. J. E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, Vol. 36, 1964, No. 8, pp. 1627–1639.

[31] Havens, T.—Bezdek, J.—Keller, J.—Popescu, M.: Clustering in Ordered Dissimilarity Data. Technical report, Univ. of Missouri 2007.

[32] Myrtveit, I.—Stensrud, E.—Olsson, U. H.: Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods. IEEE Transactions on Software Engineering, Vol. 27, 2001, pp. 999–1013.

[33] Al-Shalabi, L.—Shaaban? Z.—Kasasbeh, B.: Data Mining: A Preprocessing Engine. Journal of Computer Science, Vol. 2, 2006, No. 9, pp. 735–739.

[34] UCI Repository of Machine Learning Databases (2008), `http://www.ics.uci.edu/`.

[35] Liu, Y.—Li, Z.—Xiong, H.—Gao, X.—Wu., J.: Understanding of Internal Clustering Validation Measures. IEEE International Conference on Data Mining 2010.

[36] Zelnik, L.—Manor-Perona, P.: Self-Tuning Spectral Clustering. Advances in Neural Information Processing Systems, MIT Press 2004.

[37] Ning, H. Z.—Xu, W.— Chi, Y.—Huang, T. S.: Incremental Spectral Clustering with Application to Monitoring of Evolving Blog Communities. Proc. SIAM Int'l Conf. Data Mining 2007.

[38] Xu, W.—Liu, X.—Gong, Y.: Document Clustering Based on Non-Negative Matrix Factorization. Proc. ACM SIGIR 2003.

[39] Lovasz, L.—Plummer, M .: Matching Theory. Elsevier Science Publishers B. V. and Akadémiai Kiadó, Budapest 1986.

[40] Mirkin, B.: Clustering for Data Mining: A Data Recovery Approach. Chapman and Hall/CRC 2005.

**Puniethaa Prabhu** received her B. Sc. in computer science and Master of Computer Application from Bharathiyar University, Coimbatore, India in 1995 and 1998, respectively. Currently she is working towards her Ph. D. degree in Anna University, Chennai, India, and working as an Associate Professor with K. S. Rangasamy College of Technology (Autonomous Institution) from 1998. Her research interests include bioinformatics, data mining, automatic clustering and machine learning. She is a life time member in MISTE.

**Karuppusamy Duraiswamy** received the B. E., M. Sc. and Ph. D. degrees from the University of Madras and Anna University in 1965, 1968 and 1987, respectively. He worked as a Lecturer in the Department of Electrical Engineering in Government College of Engineering, Salem from 1968, as an Assistant Professor in Government College of Technology, Coimbatore from 1983 and as the Principal at K. S. Rangasamy College of Technology from 1995. He is currently working as a Dean in the Department of Computer Science and Engineering at K. S. Rangasamy College of Technology (Autonomous Institution). His research interests include mobile computing, soft computing, computer architecture and data mining. He is a senior member of ISTE, IEEE and CSI.