INT J COMPUT COMMUN, ISSN 1841-9836
8(2):275-293, April, 2013.

Enhanced Dark Block Extraction Method Performed Automatically to Determine the Number of Clusters in Unlabeled Data Sets

P. Prabhu, K. Duraiswamy

Puniethaa Prabhu, K. Duraiswamy
Department of Master of Computer Application
K.S. Rangasamy College of Technology
Tamil Nadu, India.
Email: spunitha156@yahoo.co.in, drkduraiswamy@yahoo.co.in

Abstract: One of the major issues in data cluster analysis is to decide the number of clusters or groups from a set of unlabeled data. In addition, the presentation of cluster should be analyzed to provide the accuracy of clustering objects. This paper propose a new method called Enhanced-Dark Block Extraction (E-DBE), which automatically identifies the number of objects groups in unlabeled datasets. The proposed algorithm relies on the available algorithm for visual assessment of cluster tendency of a dataset, by using several common signal and image processing techniques. The method includes the following steps: 1.Generating an Enhanced Visual Assessment Tendency (E-VAT) image from a dissimilarity matrix which is the input for E-DBE algorithm. 2. Processing image segmentation on E-VAT image to obtain a binary image then performs filter techniques. 3. Performing distance transformation to the filtered binary image and projecting the pixels in the main diagonal alignment of the image to figure a projection signal. 4. Smoothing the outcrop signal, computing its first-order derivative and then detecting major peaks and valleys in the resulting signal to acquire the number of clusters. E-DBE is a parameter-free algorithm to perform cluster analysis. Experiments of the method are presented on several UCI, synthetic and real world datasets.

Keywords: Enhanced DBE, Automatic clustering, Cluster tendency, Visual assessment, Reordered dissimilarity image.

1 Introduction

The major concern in data mining is to outline the observed data into knowledge structures. Clustering aims at classifying objects of a related class into their relevant categories. Partitioning the set of objects $O = (o_1, o_2, ..., o_n)$ into C self-related objects is the major process of cluster analysis. Various clustering algorithms are reported in the literature [1] and [2]. 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 discovered clusters C. This paper addresses the problem of determining whether the clusters are present by assessing of clustering tendency of clustering tendency as a prior process before clustering. Majority of the clustering algorithms need the number of clusters C as a key factor, so the quality of the resultant clusters mainly depends on the assessment of C.

Jain and Dubes [3] had discussed several statistically based informal techniques for cluster tendency assessment. Ling [4] proposed a clustering algorithm based on estimated distribution model. Cattell [5] formerly depicted pairwise dissimilarity information about a data set including n objects as an $n \times n$ image, where the objects are suitably reordered so that the resultant image is improved and is capable to emphasize the possible cluster structure in the data. The major papers in the visual representation of data dissimilarity include the contribution of [6], [7], [8] and [9]. The universal denominator in all this methodology is reordered dissimilarity image (RDI). The intensity of each pixel in the RDI represents the dissimilarity between the pair of objects denoted by the row and column of the pixel. An observer can merely calculate approximately the number of clusters C (i.e., count the number of dark blocks along the diagonal) of an RDI where the dark blocks posse's image lucidity (see Figure 1c).

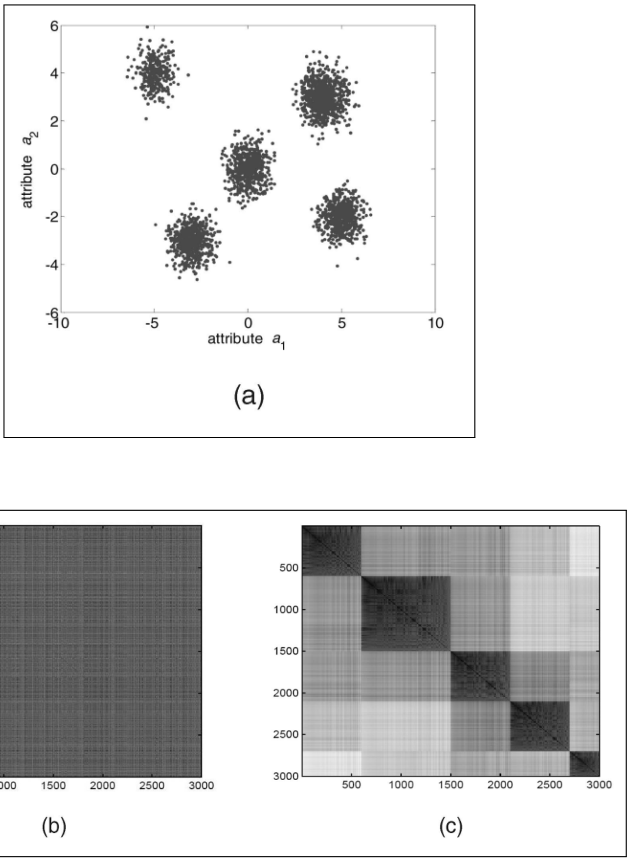


Figure 1: An example for E-VAT image. (a) Scatter plot of a 3,000 - point's data set with five clusters (b) Unordered image (c) Reordered E-VAT image $I(\bar{D})$.

Generating RDIs could be done from any of the schemes anticipated in [6], [7], [8] [9] and [11]. This paper develops a novel method to estimate automatically the number of dark blocks (seemingly also the number of possible clusters) in RDIs of unlabeled data sets. The proposed Enhanced dark block extraction (E-DBE) process combines several common images, signal processing techniques [10] and for the compactness, RDIs are generated using Enhanced Visual Assessment of Cluster Tendency (E-VAT) algorithm [11]. Later sequential image processing operations (region, segmentation, directional morphological filtering, and distance transformation) are performed to fragment the regions of interest in the RDI and then translate the filtered image into a distance-transformed image. Lastly, the altered image is projected on the diagonal axis of the RDI, which yields an one-dimensional signal from which the (potential) number of clusters can be extracted from the dataset using signal processing operations.

The rest of this paper is structured as follows: In Section 2 we present the literature description of visual approach. Section 3 reviews the enhanced VAT algorithm and Section 4 explains the procedure for Cluster Count Extraction (CCE) [12]. Section 5 analyses the dark block extraction algorithm. Section 6 describes the proposed Enhanced DBE approach. Section 7 provides results on UCI, synthetic and real world data sets for the proposed algorithm. The final section contains a short discussion on results and for future study.

2 Literature review

Some existing approaches of the post clustering cluster validity problem are reviewed before reciting the visual methods for cluster tendency assessment.

Index-based methods for cluster validity usually underline the intracluster density, intercluster division and additional factors such as geometric or statistical properties of the data are proposed in [13], [14], [15], [16], [17], [18], [19], [20] and [24]. For instance, Milligan and Cooper [13] compared 30 indices over a sequence of synthetic data sets. Above all these, Calinski and Harabasz [14] index seems to be the best which performs the ratio between the traces of the between-cluster and within-cluster scatter matrix. It is a significant noting that the validity indices are completely dependent on the data and algorithm used to find partitions.

Probabilistic indices of cluster validity attempt to validate the number of clusters found by probabilistic clustering algorithms. Guo [21] proposed a cluster number choice method for a small set of samples using a Bayesian Ying-Yang (BYY) model. Comparative studies such as [17] and [22] provided experimental comparisons of many criteria such as Akaike's Information Criterion (AIC), Minimum Description Length (MDL), and (BYY) for determining the number of clusters based on a Gaussian mixture model. A variety of statistical techniques for tendency assessment are discussed in the work of Jain and Dubes [3].

Visual methods for cluster tendency assessment for a range of data analysis problems have been extensively studied in [23]. Cattell [5] used single-linkage heuristics to rearrange the elements of small dissimilarity matrices, which were consequently hand-rendered for viewing. Floodgate and Hayes [7] offered hand-rendered pictures like Sneath's, but reordering was done computationally using single-linkage clustering. Majority of the clustering algorithm builds RDIs prior to clustering and the RDI is viewed as a visual aid to tendency assessment. This is the problem addressed by the new E-DBE algorithm, which uses the DBE algorithm of Liang [25] and E-VAT algorithm [11] to find RDIs and the number of clusters automatically.

A number of significant advantages of E-DBE over index-based or probabilistic methods are summarized as follows:

- E-DBE is a preclustering technique, i.e., it does not need the data to be clustered, nor does it locate clusters in the data. On the other hand, the consistency (and weakness) of postclustering index-based methods is entirely dependent on the clustering algorithms used to identify the partitions.
- Index-based post clustering methods regularly need clustering to be performed several times using a variety of cluster numbers and often find the top partition according to some predefined criteria. Repetitive clustering can be computationally expensive, particularly when the range of possibe values of C remains uncertain. E-DBE has no such constraint and is performed just once

3 Review of Enhanced Visual Assessment Tendency

Of the many achievable ways to obtain an RDI, apply E-VAT to generate RDIs of unlabeled data, i.e., to secure inputs to E-DBE algorithm. Let $O = (o_1, o_2, o_3 ... o_n)$ represent n objects in the data. Vectorial data have the type $F = (f_1, f_2, f_3 ... f_n), f_i \subset R_h$, where every coordinate of the vector f_i provides an attribute value of each of h features (i.e., $a_j, j = 1, 2, 3 ... h$) corresponding to an entity O_i. Constantly translate F into dissimilarities $D = [d_{ij} = ||f_i - f_j||]$, $1 \geq d_{ij} \geq 0$; $d_{ij} = d_{ji}$; $d_{ii} = 0$, for $1 \leq i,\ j \leq n$. To make the paper self-sufficient, review of reordering method E-VAT is shown in Table 1 which is proposed by [11] and an instance is shown in Figure 1.

Table 1
Enhanced-Visual Assessment Tendency Algorithm

Input

Consider the dataset as n x n dissimilarity matrix.

$$D = [d_{ij}]\text{where}1 \geq d_{ij} \geq 0; d_{ij} = d_{ji};\ d_{ii} = 0, for1 \leq i,\ j \leq n$$

Process

Step (1): Transform D to a new dissimilarity matrix R with $d_{ij} = 1 - exp(-d_{ij}/\sigma)$, where σ is a scale parameter determined from D using the algorithm of Otsu [26] automatically.

Step (2): Form an RDI image $I^{(1)}$ corresponding to R using the VAT algorithm [9].

Step (2.1): Let $I = \Phi,\ J = 1, 2, ...n$ and $P = (0, ...0)$.
Choose $(i, j) \in arg_{p_{jandq}} \in_j max\,\{d_{pq}\}$
Place $P(1) = i,\ I \leftarrow i$ and $J \leftarrow J - \{i\}$

Step (2.2): Iterate for $t = 2...n$
Select $(i, j) \in arg_{p_{iandq}} \in_j min\,\{d_{pq}\}$
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

Figure 1a shows the scatter plot of $n = 3,000$ records points in R^2, which are created from a combination of $C = 5$ bivariate normal distributions. These data points are transformed to a $3,000 \times 3,000$ dissimilarity matrix D by using distance measures for calculating distance between each pair of points. The five visually obvious clusters in Figure 1a are reflected by the five separate dark blocks along the main diagonal in Figure 1c, which is the E-VAT image of the records after reordering. On comparing with Figure 1b, which is the image of dissimilarities D in original input order, reordering is essential to expose the fundamental cluster structure of the data.

The following are some points about E-VAT:

- E-VAT algorithm is performed to determine the number of clusters prior to clustering. Even if the estimated result does not match with the true value, it provides a basis for setting the range.
- E-VAT depends merely on the input D, so a good quality D is decisive when D is a derivative of object vectors. If the input dataset is of high dimensionality nonlinearly separable, it may be improved by performing feature extraction.

4 Cluster Count Extraction (CCE) for Cluster Tendency Performance

In the following sections, the performance of E-DBE is compared with other preclustering assessment of cluster tendency techniques like DBE [25] and CCE algorithm [12]. CCE also counts dark blocks in RDIs using image transformation techniques. The major steps for this algorithm are summarized in Table 2.

Table 2
The Cluster Count Extraction (CCE) Algorithm

Input

$n \times n$ - scaled matrix of dissimilarities $D = [d_{ij}]$ and its VAT image $Image(D')$ scaled so that $max = white$ and $min = black$.

Step (1): Threshold $Image(D\prime)$ with Otsu's algorithm [26].

Step (2): Create a correlation filter ratio of size $s\prime$.

Step (3): Apply the Fast Fourier Transform (FFT) to both the segmented RDI and the filter.

Step (4): Proliferate tranformed VAT image with the composite conjugate of the transformed filter.

Step (5): Compute inverse FFT for the filtered image.

Step (6): Acquire the off-diagonal pixel values (e.g., p^{th} off-diagonal) of the back-transformed image and calculate its histogram.

Step (7): Cut the histogram at an arbitrary horizontal line $f = w$ and calculate the numeral of spikes.

Output

The number of dark blocks along the diagonal of $Image(D\prime)$ called as C (Cluster Interger).

The CCE algorithm is applicable to built RDIs by any of the methods obtainable in the literature. In this algorithm VAT [9] was used to obtain RDIs from D, but E-VAT [11] is used in the proposed E-DBE algorithm. CCE algorithm is performed based on the parameter settings suggested in [12], i.e., $s\prime = 20,\ p = 1$ and $w = 0$. Section 7 analyzes the results of CCE with DBE and proposed E-DBE on various synthetic, UCI Repository and Real-world datasets. The result in Table 5 shows that E-DBE is more consistent than CCE because CCE algorithm performs on off-diagnal pixels values of the images which show the poor performance of the method.

5 Review of Dark Block Extraction (DBE)

Liang [25] proposed the Dark Block Extraction algorithm to estimate the cluster number in unlabeled data sets. DBE algorithm counts the dark blocks along the diagonal of an RDI using basic image processing techniques. The method is summarized in Table 3.

Table 3
The Dark Block Extraction Algorithm

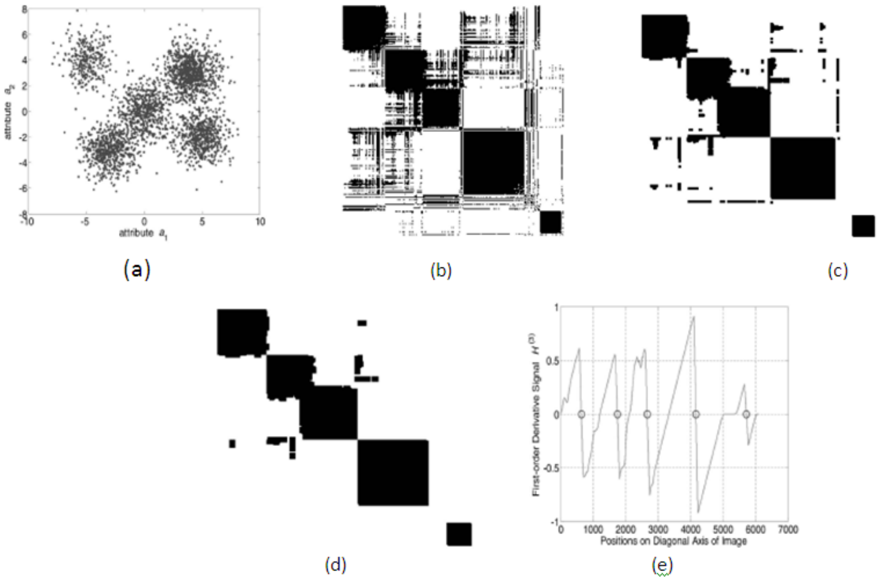


Figure 2: DBE results on five-cluster data sets. (a) Scatterplot of a data set of 3,000 points. (b) Segmented VAT image. (c) Morphological filtering of VAT image. (d) Tranformed Image (e) Signal Histogram

Input

$n \times n$ - scaled matrix of dissimilarities $D = [d_{ij}]$, the proportion of the allowed minimum cluster size of the data size n.

Step (1): Transform D to a new dissimilarity $matrixD\prime$ using σ - scale parameter determined using Otsu [26] automatically.

Step (2): Form an RDI image using VAT algorithm proposed by [9].

Step (3): Filter the image using morphological operators with directional *line* structuring elements.

Step (4): Perform a distance transform on the image to obtain a new gray-scale image.

Step (5): Project the pixel values of the image onto the main diagonal axis of the image to form a projection signal.

Step (6): Smooth the obtained signal to get the filtered signal using filter techniques.

Step (7): Find peak positions P_i and valley positions V_j in the signal.

Step (8): Select major peaks and valley by removing minor ones using filters.

Output The numbers of dark blocks (i.e., the number of major peaks) are in the RDI.

The results of dark block extraction algorithm are displayed in Figure 2. Figure 2a shows the scatter plot of 3,000 points. These points are converted to a dissimilarity matrix and perform global threshold which makes the image pixel belong to one of two classes, i.e., background or foreground.

The dissimilarity new matrix is transformed to an RDI using VAT algorithm, which is shown in Figure 2b. Then to compose the segmented image comprehensible morphological operators [25] are applied to perform binary image filtering using *line* structural element, as shown in Figure 2c. Later perform the Distance Transformation to the binary image to acquire a new gray-scale image, smooth filter techniques are applied to obtain peak and valley positions, the results are shown in Figure 2d and 2e respectively. As per Liang [25] suggestions and future developments, DBE uses simple euclidean space to calculate pairwise dissimilarities when the input records are feature vectors. The euclidean distance may not be appropriate for high dimensional or composite data. The results of DBE are not clear and the cluster extraction is not performed accurately due to VAT, datasets and thresholding. Enhanced algorithms propose a good quality dissimilarity measures for diverse types of given data sets.

6 Enhanced Dark Block Extraction (E-DBE)

The proposed algorithm extends a nearly parameter-free technique, called E-DBE, to estimate the cluster number in unlabeled data sets. E-DBE is an algorithm that counts the dark blocks along the diagonal of the RDI. The proposed method relies on E-VAT [11], Dark Block Extraction [25] and distance measures for diverse type of attributes, basic image and signal processing techniques [10] later the algorithm is summarized in Table 4.

Table 4
The Enhanced Dark Block Extraction Algorithm

Input

$n \times n$ - scaled matrix of dissimilarities $D = [d_{ij}]$ and a parameter α, the proportion of the allowed minimum cluster size of the data size n.

Step (1): Transform D to a new dissimilarity $matrixD\prime = d_{ij} = 1 - exp(-d_{ij}/\sigma)$. σ - scale parameter determined D using Otsu [26] automatically.

Step (2): Form an RDI Image(1) corresponding to $D\prime$ using E-VAT algorithm.

Step (3): Threshold the Image(1) to obtain binary Image(2) using the adaptive threshold [27] 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 [29].

Step (7): Compute the first order derivative of the Histogram(1) to obtain signal Histogram(2).

Step (8): Select major peaks and valley by removing minor ones using a filter with size α as an optimal one.

Output

The number of dark blocks (i.e., count the number of major peaks) presented in the RDI.

Prior to step-1 the datasets are preprocessed (normalized) based on the feature characteristics. Later the normalized datasets are tranformed to a dissimilarity matrix D of $n \times n$ size and the distance measure used here is the city-block distance. The matrix D is the input for constraint free algorithm E-DBE. Histogram of the original dissimilarity matrix D related to the scatter plot data set is shown in Figure 3a.

Dissimilarity Transformation and Image segmentation (Step-1): First tranform the original matrix D to a new dissimilarity matrix $D\prime$ using a *monotonic* exponential function $f(v) = 1 - exp(-v/\sigma)$ (parameter σ may be merely selected as the threshold significance obtained by Otsu's algorithm [26]), shown in Figure 3b. The resultant histogram of the dissimilarity matrix $D\prime$ is shown in Figure 3c.

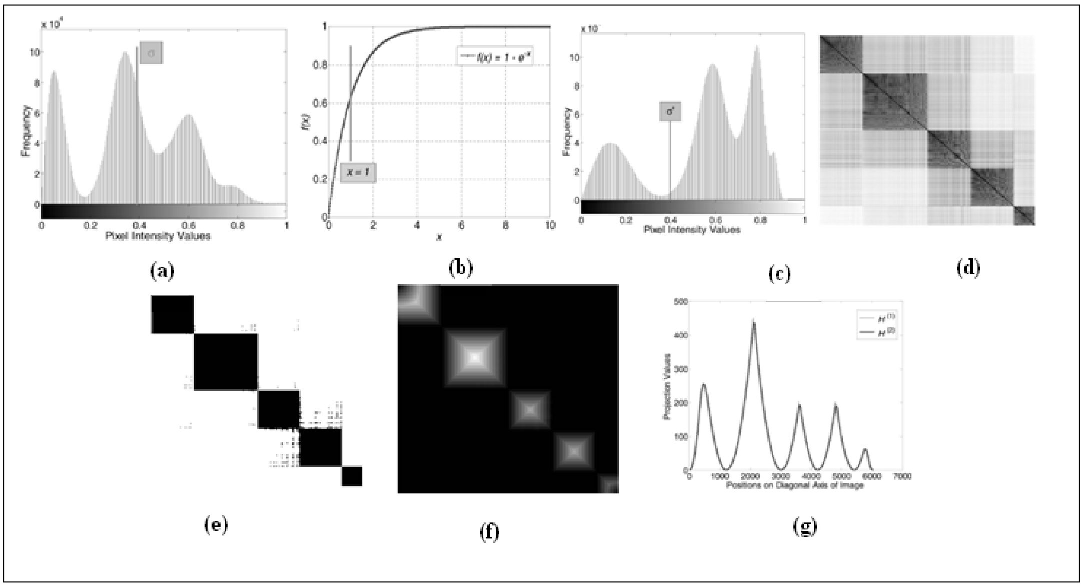


Figure 3: Sample results of E-DBE algorithm. (a) Histogram of the original dissimilarity matrix D related to the data set in Fig. 1a. (b) Monotonic transformation function, i.e., $f(v) = 1 - exp(-v/\sigma)$ (c) Histogram of $D\prime$. (d) E-VAT image ($Image^{(1)}$) (e) Segmentation of E-VAT image ($Image^{(2)}$) is obtained from binary image ($Image^{(1)}$) (f) Distance transformation on segmentated image ($Image^{(2)}$) to get a gray-scale image ($Image^{(3)}$), (g) Diagonal projection signal $Histogram^{(1)}$ from $Image^{(3)}$ and its equivalent smoothed signal $Histogram^{(2)}$.

Formation of RDI image (Step-2): In this step, E-VAT algorithm is applied to reorder the new dissimilarity matrix $D\prime$ to form an RDI image $Image^{(1)}$. E-VAT algorithm is performed to find out the figure of clusters previous to clustering. The intensity of every pixel in the RDI represents the dissimilarity among the couple of items denoted by the row and column of the pixel. RDI highlights the achievable clusters as a place *darkblocks* beside the diagonal of the image, corresponding to sets of objects with small dissimilarity. The RDI of the new dissimilarity matrix $D\prime$ is displayed in the Figure 3d and shows the formation of 5 clusters in the dataset.

Formation of binary image using adaptive threshold algorithm (Step-3): Adaptive thresholding [27] normally accepts a gray scale as input and, in the simplest execution, outputs a binary image representing the segmentation. For every pixel in the image, a threshold has to be calculated. If the pixel significance is lower than the threshold it is set to the background; otherwise it assumes the foreground value. In the proposed E-DBE adaptive or dynamic thresholding algorithm is performed which acquire best result when compared with the previous algorithm.

Adaptive threshold algorithm is used to obtain a new threshold $\sigma\prime$ to convert the binary image by $Image^{(2)}_{ij} = 1$ if $Image^{(1)}_{ij} > \sigma\prime$and $Image^{(2)}_{ij} = 0$ otherwise. The resultant binary image $Image^{(2)}$ is shown in the Figure 3e.

Distance transform of binary image to a new gray-scale image (Step-4): To organize the filtered image into an informative one showing the dark block structure information, the values of pixels that are beside or off the main diagonal axis of the image must be considered. First, execute a DT of the binary image ($Image^{(2)}$) to obtain a new gray-scale image $Image^{(3)}$ as shown in Figure 3f. A DT 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 $Image^{(2)}$. There are numerous diverse DTs depending upon which distance measures is being used to decide the distance between pixels. Euclidean distance measure is applied in this proposed algorithm. After the distance transformation, *all* pixel values of the DT $Image^{(3)}$ are projected onto the main diagonal axis to obtain a projection signal $Histogram^{(1)}$, as shown in Figure 3g. From the figure, C can be simply calculated because of the quite clear separation between major peaks in the signal $Histogram^{(1)}$.

Detection of major peaks and valleys in the projected signal (Steps 5-8): The amount of dark blocks in any RDI is equal to the number of $majorpeaks$ in the projection signal $Histogram^{(1)}$. Based on the $first - order\ derivative$ of the projection signal the cluster number C is calculated from the detection of peaks and valleys. Although the projection signal $Histogram^{(1)}$ is available, need further smoothing to reduce possible false detections due to noise in the signal. Here Savitzky-Golay smoothing filters [29] (also called digital smoothing polynomial filters or least-squares smoothing filters) are typically used to *smooth out* a noisy signal whose frequency span is large. In this algorithm, Savitzky-Golay smoothing filters perform much better than typical averaging FIR filters performed in [25], which tend to filter out a significant portion of the signal's high frequency content along with the noise. Savitzky-Golay filters are optimal in the sense that they minimize the least-squares error in fitting a polynomial to frames of noisy data. It is well recognized that the peaks and valleys of a signal usually correspond to $zero - crossing$ points in its first-order derivative, as shown in Figure 3g.

Remark - A significant issue for the E-DBE algorithm is how to successfully set the filter size α for the Savitzky-Golay filter. Actually, α is very simple to set because it reflects the minimum support threshold for the smallest cluster of importance in the data.

The novel in this algorithm is

- After preprocessing the dissimilarity matrix is transformed to a *monotonic* exponential function.
- Distance measure used in this procedure is *CityBlock* distance which gives better results
- For better performance the proposed methods uses *adaptive threshold* for segmentation
- First order derivatives are computed
- For better projection of the signals the algorithm performs *smooth*, *moving* and savitzty-golay filters.

7 Experiment results of Synthetic, UCI and real world data sets

To assess the E-DBE algorithm with its prior measures, a number of experiments on several synthetically generated data sets, UCI *Machine Learning Repository* [30] as well as real-world data set are carried. The data sets' characteristics and the results of CCE, DBE and enhanced DBE are accomplished in Table 5.

Table 5
Summary of Synthetic, UCI and Real datasets' distinctiveness and the results using CCE, DBE and Enhanced-DBE

Data set	**# Instances**	**#Clusters**	**#Each cluster**	**Attribute type**	**# Attributes**	**CCE**	**DBE**	**E-DBE**
Synthetic Datasets								
Synthetic dataset -1	1000	2	[500,500]	Integer	2	2	2	2
Synthetic dataset -2	1000	3	[500,250,250]	Integer	2	2	2	3
Synthetic dataset -3	1800	3	[300,600,900]	Integer	2	2	2	3
UCI Datasets								
Dermatology	357	6	[110,59,70, 48,51,19]	Integer	34	1	3	6
Heart	270	2	[150,120]	Integer/ Real	13	3	1	2
Hepatisis	72	2	[12,60]	Integer/ Real	20	1	1	2
Iris	150	3	[50,50,50]	Integer/ Real	5	1	2	3
Wine	178	3	[59,71,48]	Integer/ Real	13	2	2	3
Real world Datasets								
HIV	400	6	[221,144, 11,17,5,1]	Integer/ Real	19	1	3	6

7.1 Numerical examples with Synthetic Datasets

Observe the results on several synthetic datasets with multifaceted structures, in which an apparent cluster centroid for every cluster is not automatically available. Selections of synthetic datasets are based on the sets proposed in [25]. Synthetic Dataset $(S-1)$ is composed of two half-moon like patterns $(C = 2)$. The dimension of the dataset is $n = 1000$, with 500 points in each group. The upper half-moon is generated by $f_u(\phi) = 2sin(\phi) + 0.5$ randn for $\phi = [\pi/500 : \pi/500 : \pi]$, while the lower half moon is produced by $f_1(\phi) = 2sin(\phi + 0.6\pi) + 0.5$ rand for $\phi = [0.4\pi + \pi/500 : \pi/500 : 1.4\pi]$, where *randn* is a probability number drawn from a standard distribution with a zero mean and a standard deviation of one. Synthetic Dataset $(S-2)$ is generated from a grouping of two bivariate standard distributions and one half-moon like model ($C = 3$). The magnitude of the data set is $n = 1000$, including 500 points for the half-moon pattern and 250 points for each of the two Gaussian shapes. The upper half-moon is generated by $f(\phi) = 2sin(\phi) + 0.3$ randn for $\phi = [\pi/500 : \pi/500 : \pi]$, where $rand$ is a arbitrary number drawn from a regular distribution on the part interval. The two Gaussian shapes are generated by the subsequent constituent parameters: the integration proporations are $mean_1 = 0.5$ and $mean_2 = 0.5$; the mean values $\mu_1 = (0.9, 0.5)T$ and $\mu_2 = (2.1, 0.5)T$; and the covariance matrices $\sum_1 = \sum_2 = [10; 00.1]$.

Synthetic dataset$(S - 3)$ is generated from a permutation of three circles with the identical centroid but diverse radii ($C = 3$). For every circle, generate synthetic data points by $a_1(\phi) = rsin(\phi) + brandnsin(\phi)$ and $a_2(\phi) = rcos(\phi) + brandncos(\phi)$ where b is a constraint that controls the degree of overlap linking different circles, r is the radius of every circle, $\phi = [(2\pi)/p : (2\pi)/p : 2\pi]$, and p is the size of each cluster. The synthetic datasets ($S - 1, S - 2$ and $S - 3$) outcomes of E-DBE algorithm using image processing techniques are depicted in Figure 4. The E-VAT images are shown in Figure 4a, Binary E-VAT images in 4b and the first order derivative Projection Signal obtained using *smooth*, *moving* and *sgolay* are presented in Figure 4c.

7.2 Numerical examples with UCI Machine Learning Repository

Next, consider some UCI datasets which are evaluated for the performance of proposed E-DBE method. The five datasets are dermatology, heart, hepatisis, iris and wine of UCI Machine Learning Repository [30]. For each dataset, the enhanced DBE with class attribute and dimensionality reduction [28] are performed. The UCI data sets' characteristics and the consequences of E-DBE are accomplished in Table 5.

Dermatology: The main intend of this database is to determine the category of Eryhemato-Squamous Disease. They all allocate the clinical features of erythema and scaling, with very modest differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. The dataset include 357 occurrences with 34 features including class attribute. i.e., 110 for class 1, 59 for class 2, 70 for class 3, 48 for class 4, 51 for class 5 and 19 for class 6. Starting with 34-dimensional feature vectors, dataset are subjected to preprocessing, normalization and pairwise dissimilarities using the distance measures to get relational data. Later the dissimilarity matrix D is submitted to E-DBE algorithm for automatic clustering and the results are shown in Figure 5.

Heart: This dataset encloses the results of the prediction of heart attack. The dataset contains 72 instances and 13 attributes they are age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol and fasting blood sugar etc. The entire number of illustration in this data set is n=270, i.e., 150 represent absence and 120 the occurrence of heart attack. Initially with 13-dimensional feature vectors, the dataset preprocessing, normalization and pairwise dissimilarities by the distance measures are performed to acquire relational records. Afterwards, the dissimilarity matrix D is proposed to E-DBE algorithm for automatic grouping and the outcome is depticted in Figure 6. Three clusters are shown as a result of CCE algorithm and one cluster is displayed as an outcome of DBE and two clusters by E-DBE (C=2).

Hepatisis: Hepatisis is an irritation of the liver characterized by the occurrence of inflammatory cells in the tissue of the organs. This dataset contains the facts of the patient from which we evaluate whether they are alive or not. The data set holds 72 cases with 20 features (including class) they are age, sex, steroid, antiviral, fatigue and malaise etc. The total integer of instances in this data set is n=72, i.e., 12 are scrutinized as dead, and 60 are alive. The dissimilarity matrix D is submitted to E-DBE algorithm for robotic clustering. The hepatisis dataset is submitted to analyse the concertness of CCE, DBE and E-DBE algorithms results are

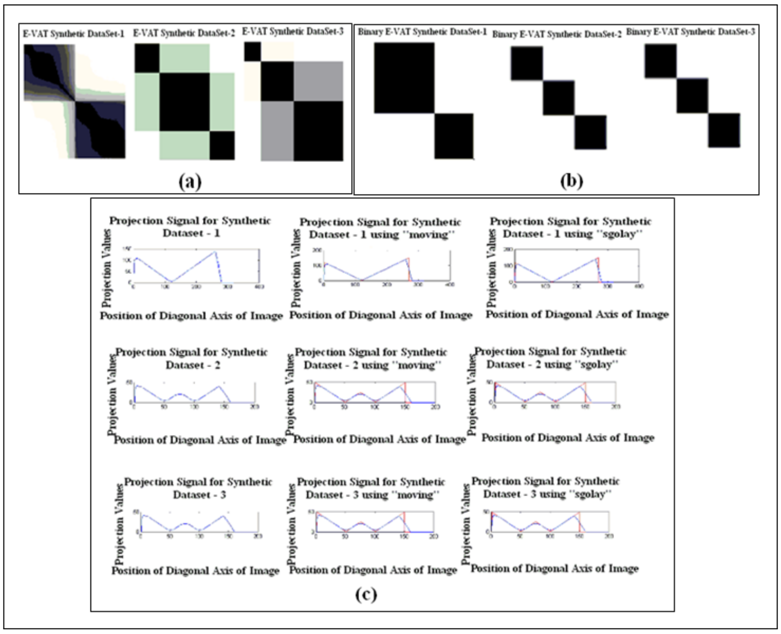
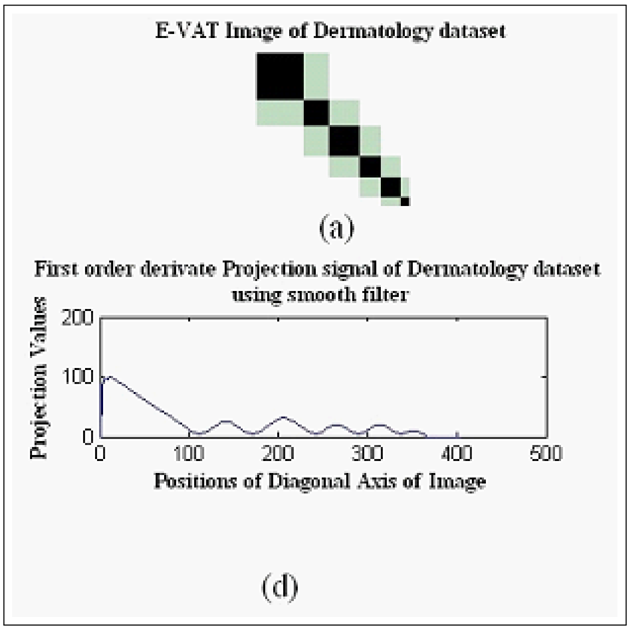


Figure 4: Results of the E-DBE algorithm on Synthetic datasets (S-1,S-2 and S-3) (a) E-VAT Images synthetic data sets (b) Binary E-VAT images (c) First order derivative Projection Signal obtained using *smooth*,*moving* and *sgolay*.



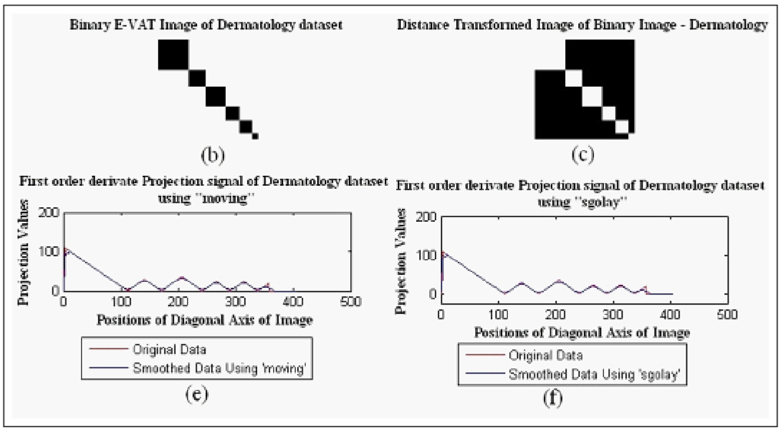


Figure 5: Results of the E-DBE algorithm on Dermatology Dataset (a) E-VAT Image of Dermatology Dataset, (b) Binary E-VAT image of Dermatology Dataset (c) Distance Transformed Image (d)First order derivative Projection Signal obtained using *smooth* (e) First order derivative Projection Signal obtained using *moving* (f) First order derivative Projection Signal obtained using *sgolay*.

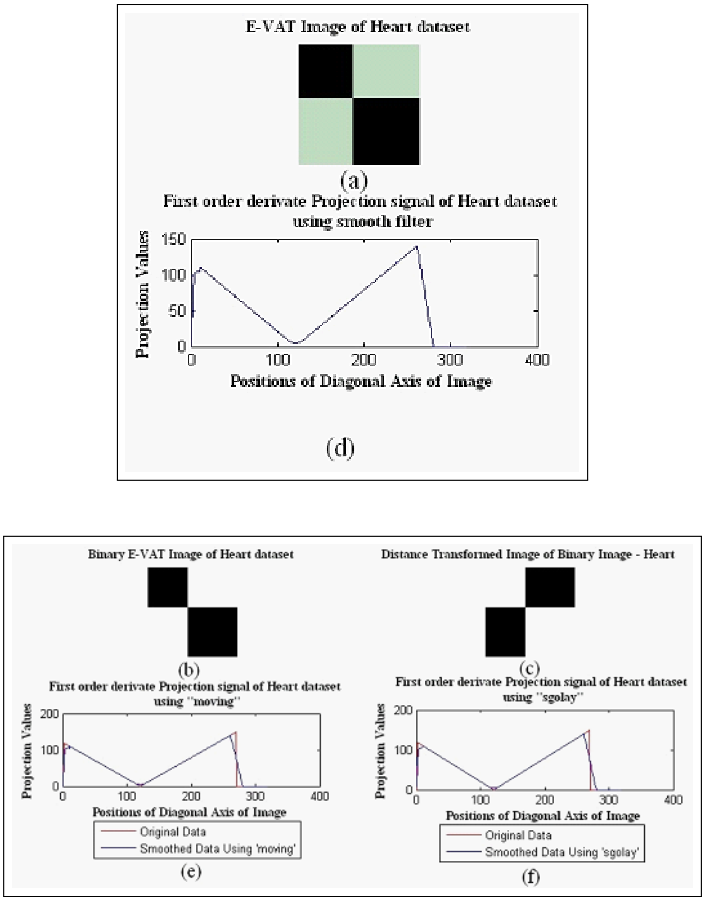


Figure 6: Results of the E-DBE algorithm on Heart Dataset (a) E-VAT Image of Heart Dataset, (b) Binary E-VAT image of Heart Dataset (c) Distance Transformed Image (d) First order derivative Projection Signal obtained using *smooth* (e) First order derivative Projection Signal obtained using *moving* (f) First order derivative Projection Signal obtained using *sgolay*.

shown in the Figure 7.

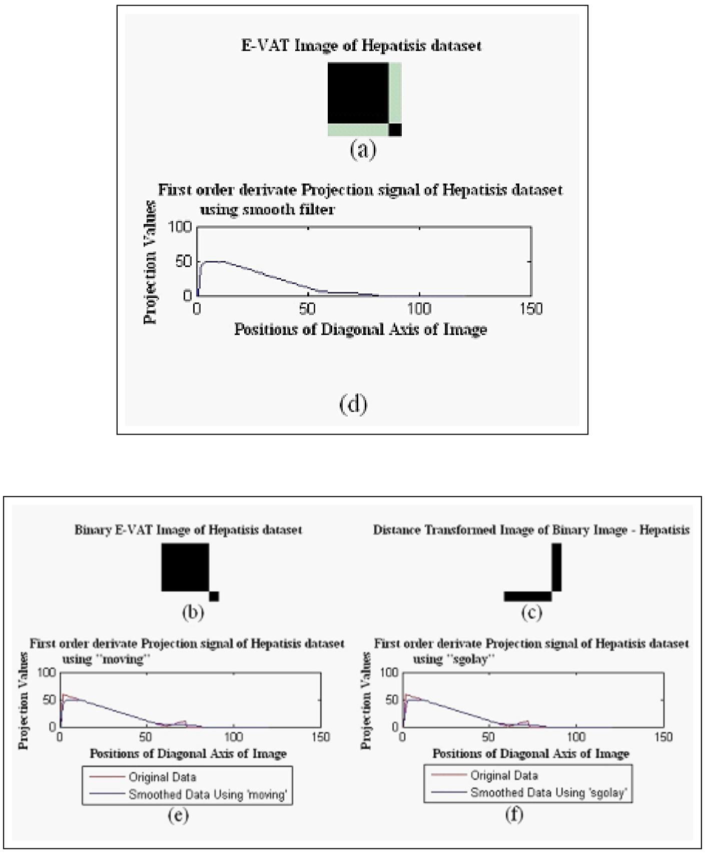


Figure 7: Results of the E-DBE algorithm on Hepatisis Dataset; (a) E-VAT Image of Hepatisis Dataset; (b) Binary E-VAT image of Hepatisis Dataset; (c) Distance Transformed Image; (d) First order derivative Projection Signal obtained using *smooth*; (e) First order derivative Projection Signal obtained using *moving*; (f) First order derivative Projection Signal obtained using *sgolay*.

Iris: This is conceivably one of the best-known databases to be found in the pattern recognition literature. The data set have 3 physical classes, 50 instances each (n=150), where each one class refers to a category of iris plant. The features of each instance consist of 4 numeric standards, consequent to sepal length, sepal width, petal length and petal width respectively. The dissimilarity matrix D is submitted to E-DBE algorithm for computerized clustering and the outcomes are shown in the Figure 8.

Wine: This data set contains the results of chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The investigation determines the quantities of 13 constituents found in each of the three brands of wines. The characteristic are respectively alcohol, malic acid, ash, magnesium, etc. The complete numeral of instances in this items are n =178, i.e., 59 for class 1, 71 for class 2 and 48 for class 3. The E-DBE results for wine data sets are shown in Figure 9.

7.3 Numerical example with Real-word Data set

The proposed method is tested on the HIV patient datasets collected from various Integrated counseling and Testing center (ICTC) and Antiretroviral (ART) centers of Tamilnadu and pondicherry. The preprocessing techniques are executed and then CCE, DBE and E-DBE

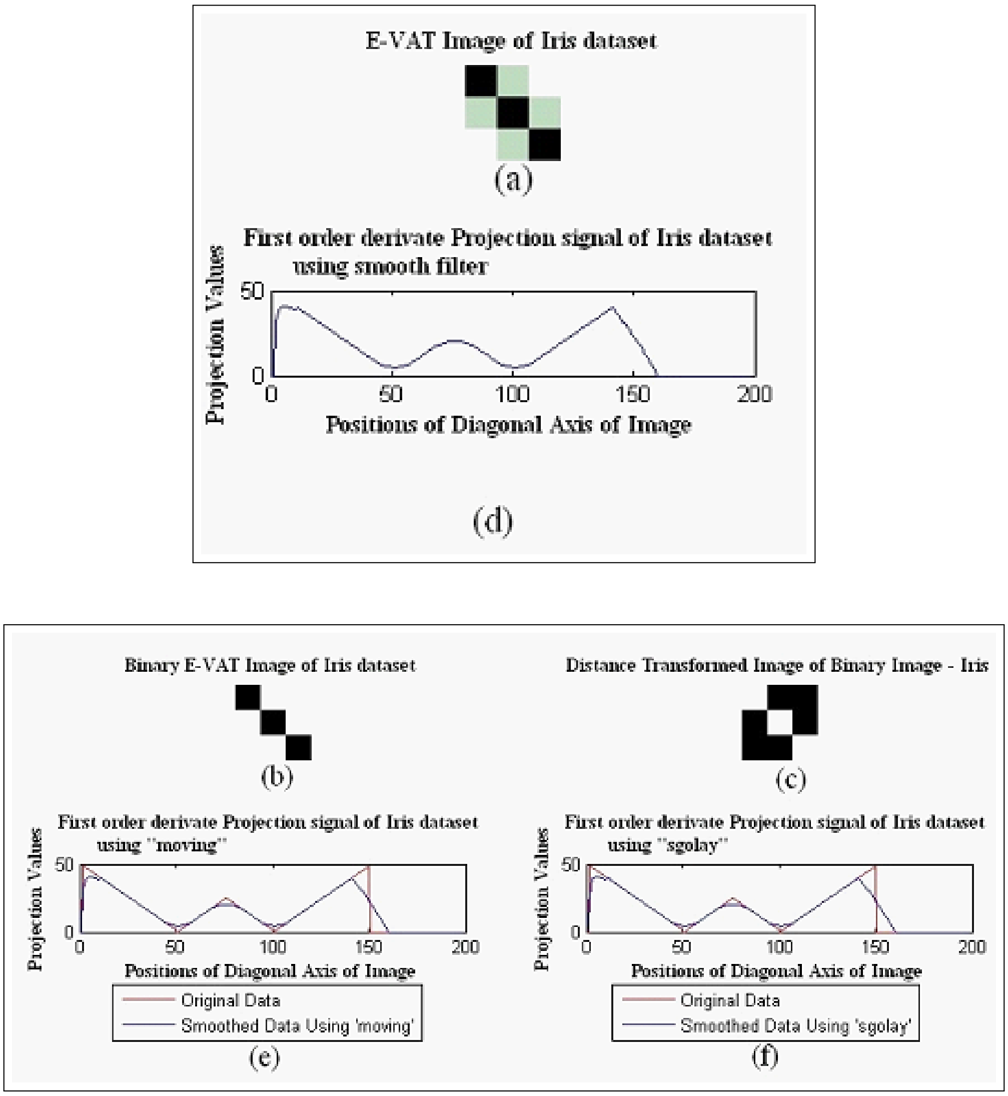
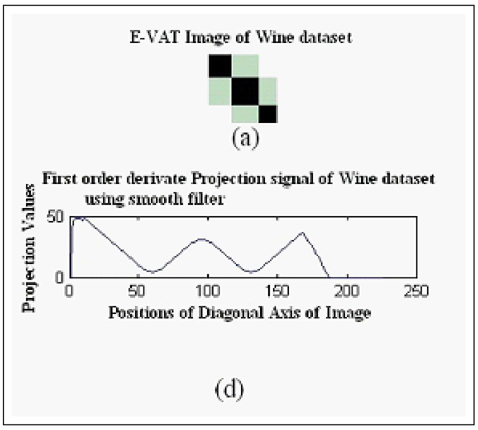


Figure 8: Results of the E-DBE algorithm on Iris Dataset; (a) E-VAT Image of Iris Dataset; (b) Binary E-VAT image of Iris Dataset; (c) Distance Transformed Image; (d) First order derivative Projection Signal obtained using *smooth*; (e) First order derivative Projection Signal obtained using *moving*; (f) First order derivative Projection Signal obtained using *sgolay*.



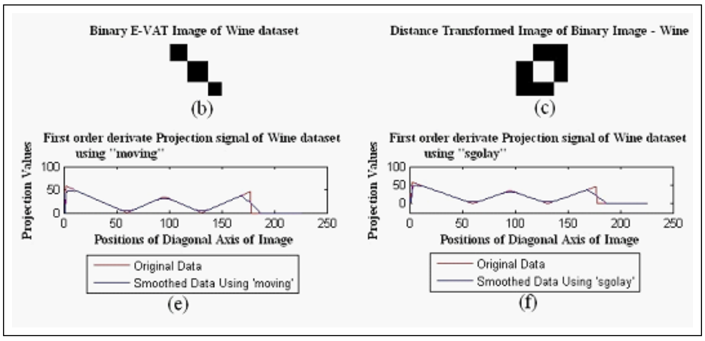


Figure 9: Results of the E-DBE algorithm on Wine Dataset (a) E-VAT Image of Wine Dataset, (b) Binary E-VAT image of Wine Dataset (c) Distance Transformed Image (d) First order derivative Projection Signal obtained using *smooth* (e) First order derivative Projection Signal obtained using *moving* (f) First order derivative Projection Signal obtained using *sgolay*.

algorithms are applied to the HIV/AIDS diagnosis dataset containing 400 objects. Table 6 shows the structure of the dataset with preprocessing depends upon the feature nature. The attributes are respectively 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 complete numeral of items in this data set is n=400, i.e., 221 for class 1, 144 for class 2, 11 for class 3, 17 for class 4, 5 for class 5 and 1 for class 6. We computed pair wise dissimilarities using the Euclidean, Hamming, Mahalanobis distance to get relational table. The E-DBE results shows the cluster count as five (C=6) which is shown in Figure 10 a better result when compared with its prior algorithm CCE and DBE.

Obj #	CA	Age	Sex	HB	WT	Treat-Drug (regimen)	. . .	CD4 Count	WBC	SGPT	TLC
1	1	25	1	14	60	1	:	500	4600	46.0	4.0
2	2	35	1	11	48	2	:	100	6400	47.0	5.0
:	1	:	:	:	:	:	:	:	:	:	:
:	1	:	:	:	:	:	:	:	:	:	:
400	2	45	0	13.5	58	1	. . .	150	3500	40.0	3.0

From the current study, the qualities of clusters are confirmed with the dark blocks on the diagonals and first order derivatives are achieved as peaks and valleys on the enhanced DBE creation. It makes certain impact of objects related to the clusters in the reversed format.

8 Discussion and conclusion

This paper examines an almost parameter-free method for automatically estimating the number of clusters in unlabeled data sets. The enhanced version of DBE algorithm works for unspecified data objects of n x n dissimilarity matrix and to estimate the feature of cluster being determined. The only user-defined constraint that must be selected ? controls the filter size for applying filtering techniques. It is comparatively easy to make a pragmatic and functional choice for ?, since it effectively specifies the smallest cardinality of a cluster relative to the number of objects in the data. The cluster number extracted by E-DBE appears to be increasingly reliable. E-DBE will perhaps reach its useful limit when the RDI created by any reordering of D is not from a well ordered dissimilarity matrix. In the proposed method distance metrics

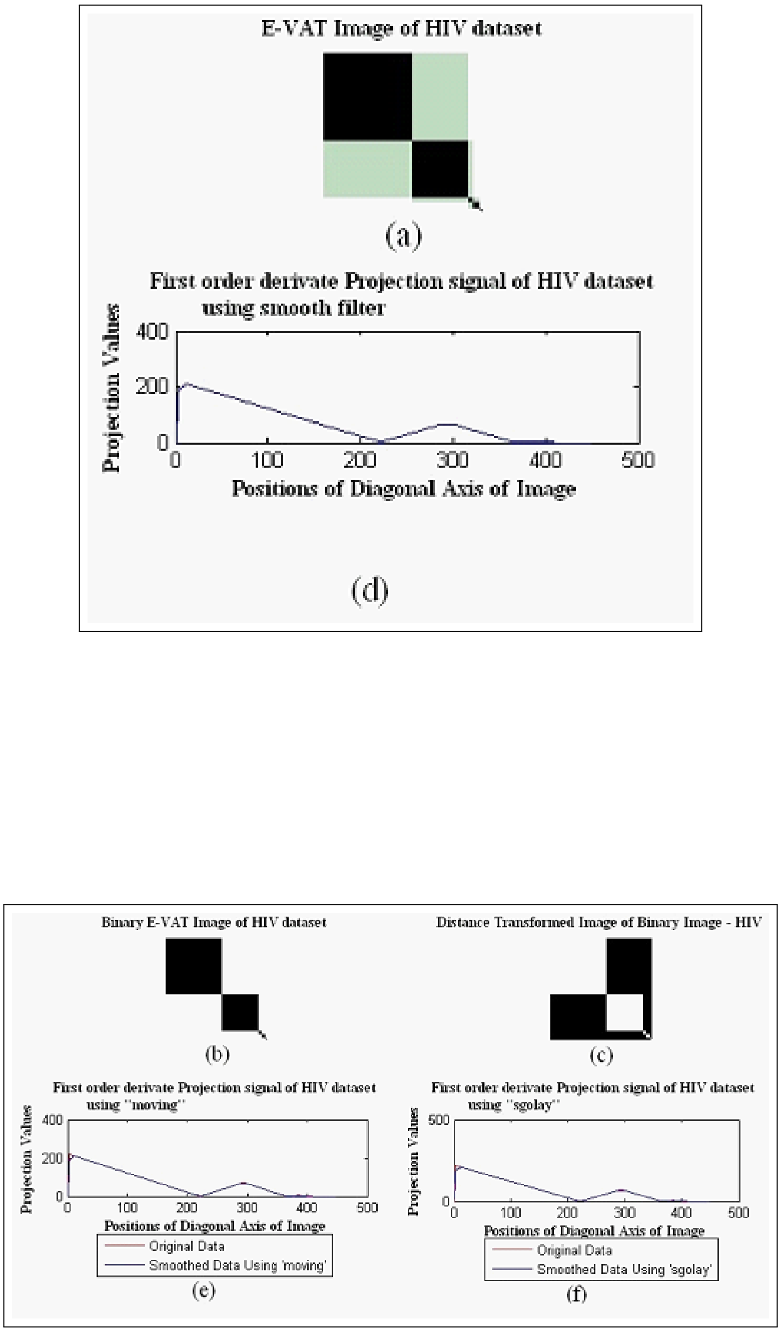


Figure 10: Results of the E-DBE algorithm on HIV- Drug Dataset (a) E-VAT Image of HIV-Drug Dataset, (b) Binary E-VAT image of HIV- Drug Dataset (c) Distance Transformed Image (d) First order derivative Projection Signal obtained using *smooth* (e) First order derivative Projection Signal obtained using *moving* (f) First order derivative Projection Signal obtained using *sgolay*.

are explored for diverse types of given data sets which yield a better cluster visualization. An achievable extension of this effort concerns the initialization of the c-means clustering algorithm for object data clustering.Future work is proposed to obtain a visual clustering algorithm based on the spectral analysis and E-VAT image and their distinctive block structured property to set the data into C clusters. By mergeing cluster tendency assessment and cluster pattern using an RDI, the proposed system can present a natural environment for visual cluster confirmation and analysis. To handle huge datasets, further propose a feasible approximate solution in a *sampling plus extension* manner to facilitate both visual cluster tendency estimation and partitioning.

Bibliography

[1] R. Xu and D. Wunsch II, Survey of Clustering Algorithms, *IEEE Trans. Neural Networks*, 16(3): 645-678,2005.

[2] Shuliang Wang , Wenyan Gan, Deyi Li and Deren Li, Data Field for Hierarchical Clustering, *Int J Data Warehousing and Mining*, 7(4): 43-63, 2011.

[3] A.K. Jain, and R.C. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[4] Ling Tan, David Taniar, Kate A. Smith, A clustering algorithm based on an estimated distribution model, *Int. J. of Business Intelligent and Data Mining*, 1(2): 229-245, 2005.

[5] R.B. Cattell, A Note on Correlation Clusters and Cluster Search Methods, *Psychometrika*, 9(3): 169-184, 1944.

[6] P. Sneath, A Computer Approach to Numerical Taxonomy, *J. General Microbiology*, 17: 201-226, 1957.

[7] G.D. Floodgate and P.R. Hayes, The Adansonian Taxonomy of Some Yellow Pigmented Marine Bacteria, *J. General Microbiology*, 30: 237-244, 1963.

[8] R.F. Ling, A Computer Generated Aid for Cluster Analysis, *Comm. ACM*, 16: 355-361, 1973.

[9] J.C. Bezdek and R. Hathaway, VAT: A Tool for Visual Assessment of (Cluster) Tendency, *Proc. Int Joint Conf. Neural Networks (IJCNN '02)*, 2225-2230, 2002.

[10] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice Hall, 2002.

[11] Puniethaa Prabhu and K.Duraiswamy, Enhanced VAT for Cluster Quality Assessment in Unlabeled Datasets, *J. of Circuits, Systems and Computers (JCSC)*, 21(1): 1-19, 2012.

[12] I. Sledge, J. Huband, and J.C. Bezdek, (Automatic) Cluster Count Extraction from Unlabeled Datasets, *Joint Proc. Fourth Int Conf. Natural Comput (ICNC) and Fifth Int Conf. Fuzzy Systems and Knowledge Discovery (FSKD)*, 2008.

[13] G. Milligan and M. Cooper, An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, 50: 159-179, 1985.

[14] R.B. Calinski and J. Harabasz, A Dendrite Method for Cluster Analysis, *Comm. in Statistics*, 3: 1-27, 1974.

[15] R. Tibshirani, G. Walther, and T. Hastie, Estimating the Number of Clusters in a Dataset via the Gap Statistics, *J. Royal Statistical Soc. B*, 63: 411-423, 2001.

[16] U. Maulik and S. Bandyopadhyay, Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(12): 1650-1654, 2002.

[17] J.C. Bezdek, W. Li, Y. Attikiouzel, and M.P. Windham, A Geometric Approach to Cluster Validity for Normal Mixtures, *Soft Computing*, 1: 166-179, 1997.

[18] J.C. Bezdek and N.R. Pal, Some New Indices of Cluster Validity, *IEEE Trans. System, Man and Cybernetics*, 28(3): 301-315, 1998.

[19] W. Wang and Y. Zhang, On Fuzzy Cluster Validity Indices, *Fuzzy Sets and Systems*, 158: 2095-2117, 2007.

[20] Decomposition Methodology for Knowledge Discovery and Data Mining, *O. Maimon and L. Rokach, eds., World Scientific*, 90-94, 2005.

[21] P. Guo, C. Chen, and M. Lyu, Cluster Number Selection for aSmall Set of Samples Using the Bayesian Ying-Yang Model, *IEEE Trans. Neural Networks*, 13(3): 757-763, 2002.

[22] X. Hu and L. Xu, A Comparative Study of Several Cluster Number Selection Criteria, *Proc. Fourth Int'l Conf. Intelligent Data Eng. and Automated Learning (IDEAL '03)*, 195-202, 2003.

[23] P.J. Rousseeuw, A Graphical Aid to the Interpretations and Validation of Cluster Analysis, *J. Comput. and Applied Math.*, 20: 53-65, 1987.

[24] Yun Sing Koh, Russel Pears and Gillian Dobbie, Automatic Item Weight Generation for Pattern Mining and its Application, *Int. J. Data Warehousing and Mining*, 7(3): 30-49, 2011.

[25] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao and James Bezdek, Automatically Determining the Number of Clusters in Unlabeled Data Sets, *IEEE Transactions on knowledge and Data Engineering*, 21(3): 335-350, 2009.

[26] N. Otsu, A Threshold Selection Method from Gray-level Histograms, *IEEE Trans. Systems, Man, and Cybernetics*, 9(1): 62-66, 1979.

[27] Mehmet Sezgin and Bulent Sankur, Survey over image thresholding techniques and quantitative performance Evaluation, *Journal of Electronic Imaging*, 13(1): 146-165, 2004.

[28] Amit Saxena and John Wang, Dimensionality Reduction with Unsupervised Feature Selection and Applying Non-Euclidean Norms for Classification Accuracy, *Int J Data Warehousing and Mining*, 6(2): 22-40, 2010.

[29] A. Savitzky and M.J.E Golay, Smoothing and differentiation of data by simplified least squares. Procedures, *Analytical Chemistry*, 36(8): 1627-1639, 1964.

[30] UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/ mlearn /ML-Repository.html.