Finding the Number of Clusters in Unlabeled Datasets using Extended Dark Block Extraction

Srinivasulu Asadi
Dept of IT
S.V.E.C, A.Rangampet,
Tirupati-517 502, India

Dr Ch D V Subba Rao
Dept of CSE
S V University,
Tirupati - 517 502, India

V Saikrishna
Dept of CSE,
S V University
Tirupati - 517 502, India

ABSTRACT
Clustering analysis is the problem of partitioning a set of objects O = {o1… on} into c self-similar subsets based on available data. In general, clustering of unlabeled data poses three major problems: 1) assessing cluster tendency, i.e., how many clusters to seek? 2) Partitioning the data into c meaningful groups, and 3) validating the c clusters that are discovered. We address the first problem, i.e., determining the number of clusters c prior to clustering. Many clustering algorithms require number of clusters as an input parameter, so the quality of the clusters mainly depends on this value. Most methods are post clustering measures of cluster validity i.e., they attempt to choose the best partition from a set of alternative partitions.

In contrast, tendency assessment attempts to estimate c before clustering occurs. Here, we represent the structure of the unlabeled data sets as a Reordered Dissimilarity Image (RDI), where pair wise dissimilarity information about a data set including 'n' objects is represented as nxn image. RDI is generated using VAT (Visual Assessment of Cluster tendency), RDI highlights potential clusters as a set of "dark blocks" along the diagonal of the image. So, number of clusters can be easily estimated using the number of dark blocks across the diagonal. We develop a new method called "Extended Dark Block Extraction (EDBE) for counting the number of clusters formed along the diagonal of the RDI. EDBE method combines several image and signal processing techniques.

General Terms: Data Mining, Image Processing, Artificial Intelligence.

Keywords — Clustering, Cluster Tendency, Reordered Dissimilarity Image, VAT, C-Means Clustering.

1. INTRODUCTION
The main Objective of our work "Estimating the number of clusters in unlabeled data sets" is to determine the number of clusters 'c' prior to clustering. Many clustering algorithms require number of clusters 'c' as an input parameter, so the quality of clusters is largely dependant on the estimation of the value 'c'. Most methods are post clustering measures of cluster validity i.e. they attempt to choose the best partition from a set of alternative partitions. In contrast, tendency assessment attempts to estimate c before clustering occurs. Our focus is on preclustering tendency assessment.

The existing technique for preclustering assessment of cluster tendency is Cluster Count Extraction (CCE). The results obtained from this are less accurate and less reliable. It does not concentrate on the perplexing and overlap issues.

Its efficiency is also doubted. Hence we are introducing a new technique in our work. Our work mainly includes two algorithms, i.e. Visual Assessment of Cluster Tendency (VAT) and Extended Dark Block Extraction (EDBE). Here, we initially concentrate on representation of structure in unlabeled data in an image format. Then for that image VAT algorithm is applied, and then for the output of VAT, we apply EDBE algorithm, there by generating the valid number of peaks (i.e. number of clusters). Pair wise dissimilarity information of a dataset including 'n' objects is depicted as an n*n image, where the objects are potentially reordered so that the resultant image is better able to highlight the potential cluster structure of the data. The intensity of each pixel in the RDI corresponds to the dissimilarity between the pair of objects addressed by the row and column of the pixel. A "useful" RDI highlights potential clusters as a set of "dark blocks" along the diagonal of the image, corresponding to sets of objects with low dissimilarity.

This dissimilarity matrix generated will be provided as input to the VAT algorithm. RDI (Reordered Dissimilarity Image) that portrays a potential cluster structure from the pair wise dissimilarity matrix of the data is created using VAT. Then, sequential image processing operations (region segmentation, directional morphological filtering, and distance transformation) are used to segment the regions of interest in the RDI and to convert the filtered image into a distance-transformed image. Finally, we project the transformed image onto the diagonal axis of the RDI, which yields a one-dimensional signal, from which we can extract the (potential) number of clusters in the data set using sequential signal processing operations like average smoothing and peak detection. The peaks and valleys are found using peak detection techniques from the projected signal. These peaks and valleys are made to satisfy certain conditions. Only the peaks which satisfy the given condition will be considered as valid peaks. The number of valid peaks provides the number of clusters that can be formed from the unlabeled data sets. The proposed method is easy to understand and implement, and encouraging results are achieved.

2. RELATED WORK
Visual methods for cluster tendency assessment for various data analysis problems have been widely studied [10], [5], [9]. For data that can be projected onto a 2D euclidean space (which are commonly depicted with a scatter plot), direct observations can provide a good insight on the value of c. Apparently, Ling [1] first automated the creation of the RDI in 1973 with an algorithm called SHADE, which was used after the application of the complete linkage hierarchical clustering scheme and served as an alternative to visual displays of hierarchically nested clusters via the standard dendrogram. Since then, there have been many studies of the best method for reordering and for the use of RDIs in clustering. Two general approaches have emerged, depending on whether the RDI is viewed before or after clustering. Most RDIs built for

viewing prior to clustering use algorithms very similar in flavor to single-linkage to reorder the input dissimilarities, and the RDI is viewed as a visual aid to tendency assessment. This is the problem addressed by our new DBE algorithm, which uses the VAT algorithm of Bezdek and Hathaway [2] to find RDIs. VAT is related but not identical to single-linkage clustering; see [11] for a detailed analysis of this aspect of VAT. Several algorithms extend VAT for related assessment problems. The bigVAT [3] and sVAT [4] offered different ways to approximate the VAT RDI for very large data sets. The coVAT [6] extended the idea of RDIs to rectangular dissimilarity data to enable tendency assessment for each of the four coclustering problems associated with such data.

2.1. Review of VAT

The visual approach for assessing cluster tendency introduced here can be used in all cases involving numerical data. It is both convenient and expected that new methods in clustering have a catchy acronym. Consequently, we call this new tool VAT (*visual assessment of tendency*). The VAT approach presents pair wise dissimilarity information about the set of objects O = {o1… on} as a square digital image with n2 pixels, after the objects are suitably reordered so that the image is better able to highlight potential cluster structure. To go further into the VAT approach requires some additional background on the types of data typically available to describe the set O = {o1… on}.

There are two common data representations of O upon which clustering can be based. When each object in O is represented by a (column) vector x in ⊂ ℜs , the set X = {x1…xn} ⊂ ℜs is called an *object data* representation of O. The VAT tool is widely applicable because it displays a reordered form of dissimilarity data, which itself can *always* be obtained from the original data for O. If the original data consists of a matrix of pair wise (symmetric) similarities S = [Sij], then dissimilarities can be obtained through several simple transformations.

For example, we can take Rij = Smax – Sij, where Smax denotes the largest similarity value. If the original data set consists of object data X = {x1,…,xn} s, then Rij can be computed as Rij = xi − x j , using any convenient norm on s .The VAT approach is applicable to virtually *all* numerical data sets

Fig. 1a is a scatter plot of n ¼ 3,000 data points in R2, These data points were converted to a 3,000 * 3,000 dissimilarity matrix D by computing the Euclidean distance between each pair of points. The five visually apparent clusters in Fig. 1a are reflected by the five distinct dark blocks along the main diagonal in Fig. 1c, which is the VAT image of the data after reordering. Compared with Fig. 1b, which is the image of dissimilarities D in original input order, we can say that reordering is necessary to reveal the underlying cluster structure of the data. The reordering method of VAT is summarized in Table 1

Fig. 1a is a scatter plot of n ¼ 3,000 data points in R2, These data points were converted to a 3,000 * 3,000 dissimilarity matrix D by computing the Euclidean distance between each pair of points. The five visually apparent clusters in Fig. 1a are reflected by the five distinct dark blocks along the main diagonal in Fig.

3. VAT ALGORITHM

Step 1) A dissimilarity matrix ‘m’ of size n*n is generated from the input dataset ‘S’, where ‘n’ is the size of ‘S’; //initialization

Step 2) set K←{1,2,3…….n}, I←J←{},P[]←{0,0,0……0};

Step 3) select (i , j) Є argmax (m_{pq}) such that (p,q) Є K and set P[1]←i; I←{i},

J←K-{i};

Step 4) for r←2,3……n

Select(i, j) Є argmin (m_{pq}) and set P[r] ←j, I←IU{j}, J←J-{j}

Next r

Step 5) Obtain the ordered dissimilarity matrix ‘R’ using the ordering array P as .

$R_{ij} = m_{p(i)p(j)}$ for 1<=i, j<=n.

Step 6) Display the Reordered Dissimilarity Image.

4. EDBE ALGORITHM

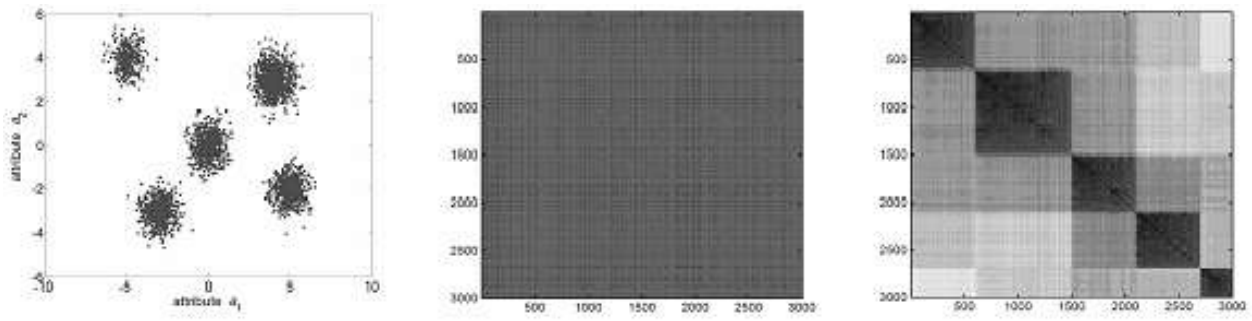
The existing system for automatically determining the number of clusters in unlabeled data sets is “cluster count extraction”.

Because of its limitations like perplexing, and its inability in histogram overlapping, we are moving on to a new technique. The proposed system is “Extended Dark Block Extraction”, which is nearly a parameter free method developed to automatically determine the number of clusters in unlabeled datasets. In short, EDBE is an algorithm that counts the dark blocks along the diagonal of a RDI.

EDBE algorithm mainly includes four major steps:

- Dissimilarity Transformation and Image segmentation.
- Directional Morphological filtering of binary image.
- Distance transform and diagonal projection of filtered image.
- Detection of major peaks and valleys in the projected signal

4.1 Dissimilarity transformation and Image



(a) Scatter plot of a 3,000 - point in data set with five cluster

(b) Unordered image

(c) Recordered V A T image I(D')

Segmentation (Steps 1-3):

Because information about possible cluster structure in the data is embodied in the dark blocks in the RDI, an important preprocessing step is image thresholding to extract the regions of interest. Choosing a threshold 'α' around the first mode is thus ideal for image segmentation. Otsu's algorithm [7], which maximizes the between-class variance, has been widely used in image processing for automatically choosing a global threshold.

$f(t) = 1 - \exp(-t / \alpha)$

EDBE ALGORITHM

Step 1) Find the threshold value 'α' from 'm' using otsu's algorithm.
Step 2) Transform 'm' in to new dissimilarity matrix 'm1' with $m1_{ij} = 1 - \exp(-m/\alpha)$
Step 3) Form an RDI image 'I^1' using the previous module.
Step 4) Threshold 'I^1' to obtain a binary image 'I^2' using algorithm of otsu.
Step 5) Filter 'I^2'using morphological operations to obtain a filtered binary image 'I^3'.
Step 6) Perform a distance transform on 'I^3' to obtain a gray scale image 'I^4' and scale the pixel values to [0, 1].
Step 7) Project the pixel values of the image on to the main diagonal axis of 'I^4' to form a projection signal 'H^1'
Step 8) Smooth the signal 'H^1' to obtain the filtered signal 'H^2' by an average filter.
Step 9) Compute the first order derivative of 'H^2' to obtain 'H^3'.
Step 10) Find peak position 'p^i' and valley positions 'v^j' in 'H^3'.
Step 11) Select valid peaks by considering some conditions. Number of valid peaks gives number of clusters.
Step 12) Put the number of clusters into C-Means Clustering Algorithm and gives very good accuracy.

This does not affect the reordering by VAT but changes the histogram of dissimilarities. From the histogram of D0, we use Otsu's algorithm again to obtain a new threshold to convert the VAT image shown in fig 2a into a binary image shown in fig 2b by

$I_{ij}^2 = 1$, if $I_{ij}^2 > \alpha$
$I_{ij}^{2=0,}$ otherwise.

It can be seen that the segmentation result after transformation is far better than that before transformation.

Directional morphological filtering of binary image (Step 4):

To make the segmented image clearer, especially for the cases in which the degree of overlap between clusters is large, we use morphological operations [8] to perform binary image filtering. Morphological filtering is one type of processing in which the spatial form or structure of objects within an image is modified. Dilation and erosion are two fundamental morphological operations. The former usually causes objects to grow in size, while the latter causes objects to shrink. The morphologically filtered image is as shown in the fig 3a

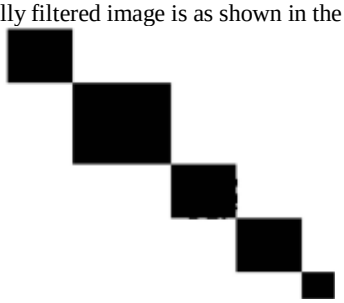


Fig 3a: Morphologically filtered Image

4.2. Distance transform and diagonal projection of image (Steps 5-6):

In order to convert the morphologically filtered image into an informative one that clearly shows the dark block structure information; we need to consider the values of pixels that are along or off the main diagonal axis of the image. First, we perform a DT of the binary image to obtain a new gray-scale image as shown in the fig 3b A DT is a form of representation 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 nearest nonzero pixel in the binary image.3

There are several different DTs depending upon which distance metric is being used to determine the distance between pixels. We use the Euclidean distance. After the DT, we project "all" pixel values of the DT image onto the main diagonal axis to obtain a projection signal as shown in the figure 3c.

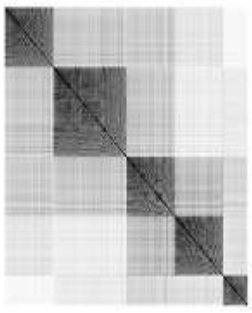


Fig 2a: VAT image of I(D')

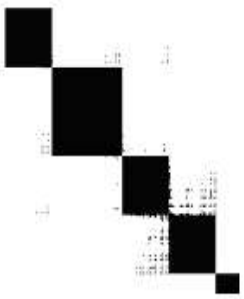


Fig 2b: Segmented image of I(D')

Fig 2c: Segmented image before transformation

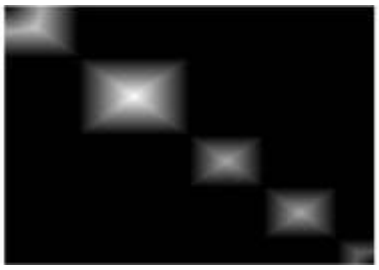


Fig3b.Distance Transformed Image(I^4)

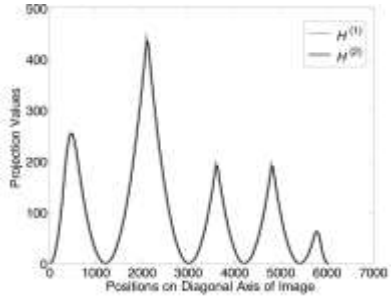


Fig 3c. Diagonal projection signal from (I^4)

4.3. Detection of major peaks and valleys in the projected signal (Steps 7-10):

The number of dark blocks in any RDI is equivalent to the number of "major peaks" in the projection signal H^1. We perform the detection of peaks and valleys to estimate the (cluster) number c, based on the "first-order derivative" of the projection signal. Although the projection signal H1 seems to be very smooth, we require further smoothing to reduce possible false detections due to noise in the signal. Here, we use a simple average filter 'h' to filter the projection signal, i.e., $H^{(2)} = \mathbf{h} * \mathbf{H}^{(1)}$, where '*' means linear convolution (see Fig. 3c), and the average filter h has length l2= 2*α*n.

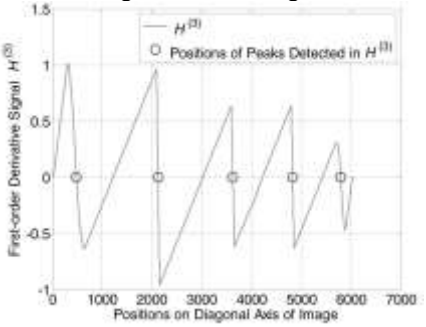


Fig 3d. First Order derivative signal

After that, the process of peak and valley detection is performed in a "from-rough-to-fine" manner. It is well known that the peaks and valleys of a signal usually correspond to "zero-crossing" points in its first-order derivative as shown in the fig.3d. Accordingly, we can find the initial sets of peaks pi and valleys vj by finding the corresponding from-positive-to-negative zero-crossing points and from-negative-to-positive zero-crossing points. To further remove minor false peaks, we use a size filter to remove relatively small valleys by validating the width between each two neighboring valleys. That is, the peak pi within the two neighboring valleys will be kept as a meaningful major peak

if $\mathbf{V_{(k+1)} - V_{(k)} > l_3}$

$\mathbf{V_{(k)} < P_{(i)} < V_{(k+1)}}$, where $\mathbf{l3=2\alpha n}$

Finally, we determine the number of dark blocks in the RDI (and, hopefully, the number of clusters c in the unlabeled data) as the number of resulting major peaks.

5. CONCLUSION

This paper investigates a nearly parameter-free method for automatically estimating the number of clusters in unlabeled data sets. The only user-defined parameter that must be chosen α controls the filter size. It is relatively easy to make a realistic (and useful) choice for α, since it essentially specifies the smallest cardinality of a cluster relative to the number of objects in the data. Cluster number should be EDBE will probably reach its useful limit when the RDI formed by any reordering of D is not from a well structured dissimilarity matrix. In our experiments, we used the simple euclidean distance to compute pair wise dissimilarities when the input data are feature vectors. The euclidean distance may not be suitable for high dimensional or complex data valleys (such as wavelet-based multiresolution analysis). EDBE provides an initial estimation of the cluster number, thus avoiding the requirement of repeatedly running a clustering algorithm multiple times over a wide range of c in an attempt to find useful clusters. In this way, EDBE compares favorably to post clustering validation methods in computational efficiency. It is noted that EDBE does not eliminate the need for cluster validity, but it simply improves the probability of success. A possible extension of this work concerns the initialization of the fuzzy post clustering algorithm for object data clustering. It should not be too hard to find an approximate center sample for each meaningful cluster from any well structured RDI.

6. REFERENCES

[1] R.F. Ling, Comm. ACM, vol. 16, pp. 355-361, 1973, "A Computer Generated Aid for Cluster Analysis,"

[2] J.C. Bezdek and R. Hathaway," Proc. Int'l Joint Conf. Neural Networks (IJCNN '02), pp. 2225-2230, 2002,

[3] J. Huband, J.C. Bezdek, and R. Hathaway, Pattern Recognition, vol. 38, no. 11, pp. 1875-1886, 2005, "bigVAT: Visual Assessment of Cluster Tendency for Large Data Sets".

[4] R. Hathaway, J.C. Bezdek, and J. Huband, Pattern Recognition, vol. 39, pp. 1315-1324, 2006, "Scalable Visual Assessment of Cluster Tendency".

[5] W.S. Cleveland, Visualizing Data. Hobart Press, 1993. [6] J.C. Bezdek, R.J. Hathaway, and J. Huband, IEEE Trans. Fuzzy Systems, vol. 15, no. 5, pp. 890-903, 2007, "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices".

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

[7] I. Dhillon, D. Modha, and W. Spangler, Proc. 30th Symp. Interface: Computing Science and Statistics, 1998, "Visualizing Class Structure of Multidimensional Data".

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

[9] T. Tran-Luu, PhD dissertation, Univ. of Maryland, College Park, 1996, "Mathematical Concepts and Novel Heuristic Methods for Data Clustering and Visualization".

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

[11] J. Huband, J.C. Bezdek, and R. Hathaway, Pattern Recognition, vol. 38, no. 11, pp. 1875-1886, 2005, "bigVAT: Visual Assessment of Cluster Tendency for Large Data Sets".

[12] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James Bezdek, Fellow, IEEE-MARCH 2009, Automatically Determining the Number of Clusters in Unlabeled Data Sets.