ISSN:0975-9646

A Comparative study of Clustering in Unlabelled Datasets Using Extended Dark Block Extraction and Extended Cluster Count Extraction

Srinivasulu Asadi, Dr.Ch.D.V.Subba Rao, O.Obulesu and P.Sunil Kumar Reddy

ABSTRACT: One of the major problems in cluster analysis is the determination of the number of clusters in unlabeled data prior to clustering. In this paper, we implement a new method for determining the number of clusters called Extended Dark Block Extraction (EDBE), which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) of a data set. Its basic steps include 1) Generating a VAT image of an input dissimilarity matrix, 2) Performing image segmentation on the VAT image to obtain a binary image, followed by directional morphological filtering, 3)Applying a distance transform to the filtered binary image and projecting the pixel values onto the main diagonal axis of the image to form a projection signal, 4) Smoothing the projection signal, computing its First-order derivative and then detecting major peaks and valleys in the resulting signal to decide the number of clusters, and 5)The C-Means algorithm is applied to the major peaks. We also implement the Extended Cluster Count Extraction (ECCE), which uses VAT and the combination of several image processing techniques. In both the methods we use Reordered Dissimilarity Image (RDI), which highlights potential clusters as a set of "Dark blocks" along the diagonal of the image, corresponding to sets of objects with low dissimilarity, which is implemented using VAT algorithm. This paper develops a new method for automatically estimating the number of dark blocks in RDI's unlabelled data sets and compares the two methods EDBE and ECCE for determining the number of clusters in unlabelled data sets.

Keywords: Clustering, Cluster tendencies, reordered dissimilarly image, VAT, ECCE, EDBE, FFT, Reordered Dissimilarity Image (RDI), C-Means.

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 dependent 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 Extended Cluster Count Extraction (ECCE). 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 thereby 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 co-clustering 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 s , the set X = {x1…xn} □ □s 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.

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

Step2)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.

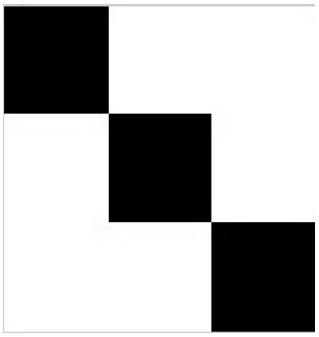

Fig 1(a): VAT Image

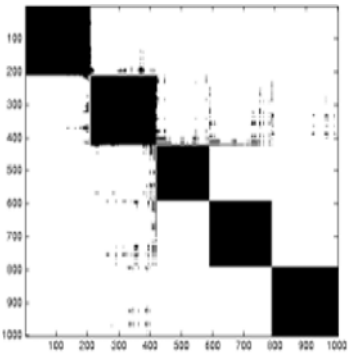

Fig 1(b): Segmented VAT image

4. ECCE ALGORITHM

The existing technique implemented for automatically estimating the number of clusters in unlabeled data sets is “EXTENDED CLUSTER COUNT EXTRACTION (ECCE)”.

Extended Cluster Count Extraction Technique

Input: n by n VAT Image, scaled so that max=white and min=black.

1. Threshold the RDI image with Otsu’s algorithm.
2. Choose a correlation filter ratio of size s.

3. Apply the Fast Fourier Transform (FFT) to both the segmented RDI and the filter.
4. Multiply the transformed RDI with the complex conjugate of the transformed filter.
5. Compute the inverse FFT for the filtered image.
6. Take the off-diagonal pixel values (e.g., q^{th} off diagonal) of the back-transformed image and compute its histogram.
7. Cut the histogram at an arbitrary horizontal line y ¼ v and count the number of spikes.
8. Perform the C-Means algorithm to the valid peaks.

Output: Integer as an estimate of number of clusters and blocks in C-Means algorithm.

In the ECCE algorithm also, we can use VAT for generating Reordered Dissimilarity Image (RDI). ECCE also counts the number of dark blocks using a combination of several image processing techniques. ECCE uses Fourier transforms, complex conjugates, Inverse Fourier transforms and many complex techniques for smoothing and filtering, where we use very simple techniques in our Extended Dark Block Extraction (EDBE) algorithm. EDBE overcomes the perplexing problem in ECCE of where to cut the histogram. For example, cutting the ECCE histogram at y ¼ 100 results in the estimate ðc ¼ 3Þ. Both algorithms require the user to specify a filter size parameter, but the rationale for the choice in EDBE is much clearer and is tied to a property of the clusters and not the RDI. In addition, the positions of peaks and valleys in EDBE implicitly correspond to centers and ranges of sub blocks (or clusters). It is hard to see similar phenomena from the ECCE histograms.

4.1 Correlation Filter Generation (Step 1 and 2):

Correlation filter is computed to filter the VAT image which is done by multiplying the complex conjugate of the filter to the VAT image after applying the Fast Fourier Transform. The Correlation Filter is calculated by taking the input as filter ratio and comparing it with the size of the input data set. The correlation Filter thus formed is a matrix of 1's and 0's after applying filter to the input and filter size ratio.

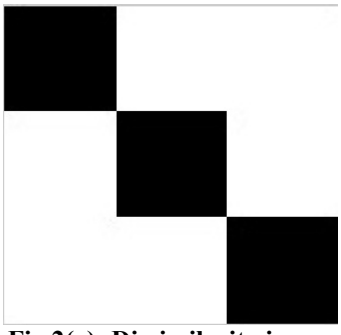

Fig 2(a): Dissimilarity image

4.2 Applying FFT to the VAT image and multiplying with the conjugate filter (Steps 3 and 4):

To begin the correlation process, both the VAT image and the corresponding detection filter are first transformed from the spatial domain to the frequency domain via the Fast Fourier Transform (FFT). Once the image is converted to the frequency domain, correlation is done by the multiplying the transformed image, with the complex conjugate of the transformed filter.

Fig 2(b): FFT VAT image

4.3 Applying Inverse FFT to the Filtered image (Step 5):

Once correlation between the segmented VAT image and filter takes place inverse Fast Fourier Transform is performed. The IFFT returns the inverse discrete Fourier transform (DFT) of vector X, computed with a fast Fourier transform (FFT) algorithm. If X is a matrix, IFFT returns the inverse DFT of each column of the matrix.

Dissimilarities, i.e. difference between every element in the matrix, with every other element is calculated and placed accordingly in the dissimilarity matrix.

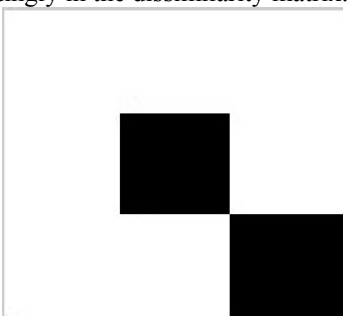

Fig 2(c): FFT Filter image

4.4 Histogram of off-diagonal pixel values of the back-transformed image (Steps 6 and 7):

Once correlation between the segmented VAT image and filter takes place, and the back-transform of the correlated image is computed, the off-diagonal values of the image are used to generate a histogram with an arbitrary number of approximately Gaussian regions that denote the preliminary number of clusters detected. Taking the set of data for some arbitrary horizontal location in the computed histogram, which will be at y=0 for the inclusion of singletons, the cluster assessment of the VAT image can be automated, with the number of clusters for that dataset returned by counting each con-tinuous distribution

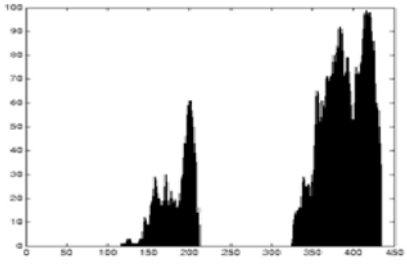

Fig 2(d): Histogram of image

4.4 C-Means Algorithm (Steps 8):

The C-Means algorithm is applied to the number of clusters from the above steps to get the centroid values and the fuzzy c-means values for each and every given data. After doing the pre-clustering to the data in VAT image, we do post-clustering to the data using C-Means algorithm. The C-Means algorithm improves the accuracy in clustering the input data sets.

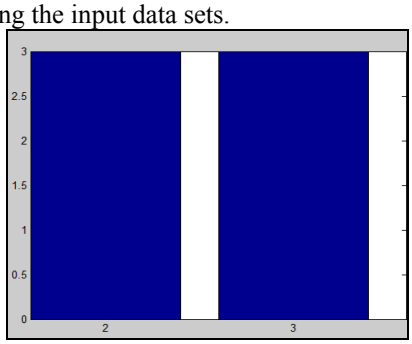

Fig 3(a): Bar Chart

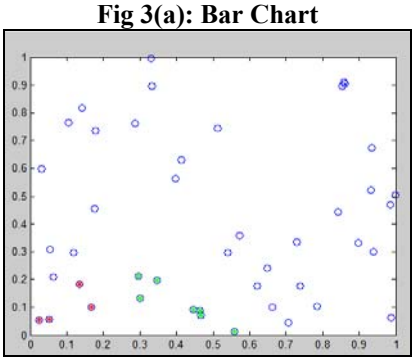

Fig 3(b): C-Means image

4.5 ECCE algorithm:

Step 1) Find the threshold value 'α' from 'm' using Otsu's algorithm

Step 2) Generate the correlation filter ratio of size s.

Step 3) Apply the FFT to the segmented VAT image and the filter.

Step 4) Multiply transformed VAT image with the complex conjugate of the transformed filter.

Step 5) Compute inverse FFT of the filtered image.

Step 6) Compute histogram of off-diagonal pixel values of the back-transformed image.

Step 7) Cut the histogram at an arbitrary horizontal line (usually 0), and count the number of spikes.

Step 8) Perform the C-Means algorithm to the valid peaks.

4.6 Disadvantages of ECCE:

EDBE is much more reliable than ECCE. The results of EDBE are more accurate than ECCE.

Mainly, in ECCE we have perplexing problem, i.e. where to cut the histogram? So that spikes will be formed and then we can count the number of spikes, there by getting the number of clusters. Estimating where to histogram itself is a problem. This problem is resolved in EDBE algorithm.ECCE fails when contiguous regions in the correlated histograms overlap. When two regions in the histogram overlap, ECCE cannot provide a solution to deal with this criterion, where as EDBE deals with the cases of overlapping regions.In ECCE, Fourier and Inverse Fourier transforms are used for smoothing, which is a very complex process. To overcome the disadvantages of the existing system, we are implementing a new technique called Extended Dark Block Extraction (EDBE).

5. 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

5.1 Dissimilarity transformation and Image 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.

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.

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

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

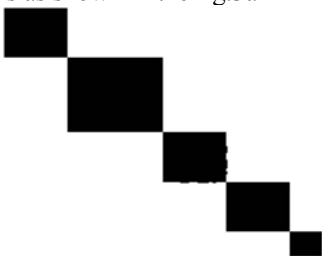

Fig 4(a): Morphologically filtered Image

5.2. 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 Distance Transform 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 fig.3b.

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 fig. 3c.

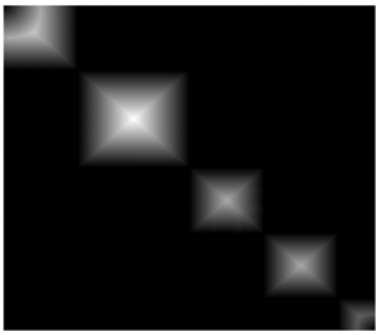

Fig 4(b) Distance Transformed Image(I^4)

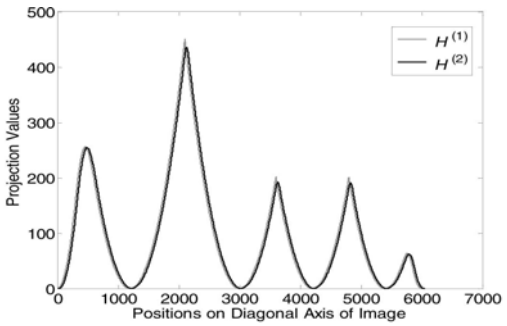

Fig 4(c): Diagonal projection signal from (I^4)

5.3. 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., $\mathbf{H^{(2)} = h * H^{(1),}}$ where '*' means linear convolution (see Fig. 3c), and the average filter h has length l2= 2*α*n.

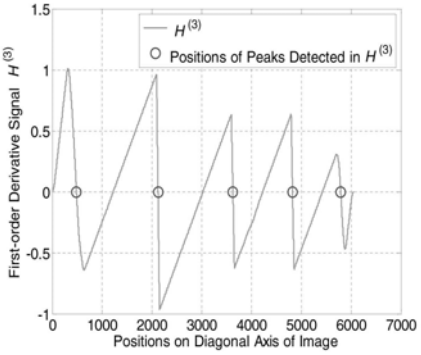

Fig 4(d): First Order derivative signal

Here is the first differential signal generated for removing any further noise using another linear filter.

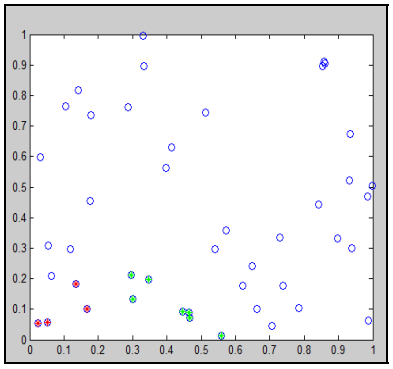

Fig 4(e): C-Means image

Here is the figure after applying c-means algorithm. We do post clustering using the C-Means Algorithm.

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

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

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.

6. CONCLUSION

This paper automatically estimates the number of clusters in unlabeled data sets. Mainly in this paper, we compare EDBE with ECCE .EDBE has no demerits and is performed well for all types of numerical unlabelled data sets. EDBE will probably reach its useful limit when the RDI is formed by any reordering of D is not from a well structured dissimilarity matrix. This method is nearly parameter-free for automatically estimating the number of clusters in unlabeled data sets. EDBE is much more reliable than ECCE. The results of EDBE are more accurate than ECCE. Mainly, in ECCE we have perplexing problem, i.e. where to cut the histogram? So that spikes will be formed and then we can count the number of spikes, there by getting the number of clusters. Estimating where to cut the histogram itself is a problem. This problem is resolved in EDBE algorithm.ECCE fails when contiguous regions in the correlated histograms overlap. When two regions in the histogram overlap, ECCE can not provide a solution to deal with this criterion, where as EDBE deals with the cases of overlapping regions. In ECCE, Fourier and Inverse Fourier transforms are used for smoothing, which is a very complex process. To overcome the disadvantages of the existing system, we are implementing a new technique called Extended Dark Block Extraction (EDBE). 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. The valid peaks are given as input to c-means to get the accurate output of number of clusters.

The possible extension of this work concerns the initialization of the fuzzy post clustering algorithm (FPCA) 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.

7. REFERENCES

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

[2] R.C. Gonzalez and R.E. Woods, Digital Image Processing. Prentic Hall, 2002.

[3] P. Soille, Morphological Image Analysis: Principles and Applications.Springer, 1999.

[4] N. Otsu, "A Threshold Selection Method from Gray-level Histograms," IEEE Trans Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, 1979.

[5] T.C. Havens, J.C. Bezdek, J.M. Keller, M. Popescu, and J.M. Huband, "Is VAT Really Single Linkage in Disguise?" Pattern Recognition Letters, 2008, in review.

[6] I. Sledge, J. Huband, and J.C. Bezdek, "(Automatic) Extended Cluster Count Extraction from Unlabeled Datasets," Joint Proc. Fourth Int'l Conf. Natural Computation (ICNC) and Fifth Int'l Conf. Fuzzy Systems and Knowledge Discovery (FSKD), 2008.

[8] 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.

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

[10] R.F. Ling, Comm. ACM, vol. 16, pp. 355-361, 1973, ―AComputer Generated Aid for Cluster AnalysisI. Dhillon, D. Modha, and W. Spangler, Proc.

[11] W.S. Cleveland, Visualizing Data. Hobart Press, 1993.

[12] P. Guo, C. Chen, and M. Lyu, "Cluster Number Selection for a Small Set of Sample Using the Bayesian Ying-Yang Model," IEEETrans. Neural Networks, vol. 13, no. 3, pp. 757-763, 2002.

[13] P. Soille, Morphological Image Analysis: Principles andApplications.Springer, 1999.

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

[15] P.J. Rousseeuw, "A Graphical Aid to the Interpretations and ValidationCluster Analysis," J. Computational and Applied Math., vol. 20, pp. 53-65, 1987.

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

[17] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12,pp. 1650-1654, Dec. 2002.

[18] R.B. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," Comm. in Statistics, vol. 3, pp. 1-27, 1974.

[19] M. Windham and A. Cutler, "Information Ratios for Validating Mixture Analysis," J. Am. Statistical Assoc., vol. 87, pp. 1188-1192, 1992.

[20] P. Gru¨ nwald, P. Kontkanen, P. Myllyma¨ki, T. Silander, and H. Tirri, "Minimum Encoding Approaches for Predictive Modeling," Proc. 14th Int'l Conf. Uncertainty in Artificial Intelligence (UAI '98), pp. 183-192, 1998.
[21] J.W. Turkey, Exploratory Data Analysis. Addison-Wesley, 1997.
[22] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Dataset via the Gap Statistics," J. Royal StatisticalSoc. B, vol. 63, pp. 411-423, 2001.
[23] UCI Repository of Machine Learning Databases, http://www.ics, uci.edu/~mlearn/MLRepository.html, 2008.
[24] I. Sledge, J. Huband, and J.C. Bezdek, "(Automatic) Cluster Count Extraction from Unlabeled Datasets," Joint Proc. Fourth Int'l Conf. Natural Computation (ICNC) and Fifth Int'l Conf. Fuzzy Systems and Knowledge Discovery (FSKD), 2008.
[25] L. Wang and D. Suter, "Visual Learning and RecognitionofSequential Data Manifolds with Applications to Human MovementAnalysis," Computer Vision and Image Understanding, 2007.
[26]A. Juan and E. Vidal, "Fast K-Means-like Clustering in MetricSpace," Pattern Recognition Letters, vol. 15, no. 1, pp. 19-25, 1994.
[27] Decomposition Methodology for Knowledge Discovery and Data Mining, O. Maimon and L. Rokach, eds., pp. 90-94. World Scientific, 2005.
[28] W. McCormick, P. Schweitzer, and T. White, "Problem Decomposition and Data Reorganization by a Cluster Technique,"Operations Research, vol. 20, no. 5, pp. 993-1009, 1972. 29] Statistical Pattern Recognition. A. Webb, ed., pp. 345-357. John Wiley & Sons, 2002.
[30] A. Gordon, Classification, second ed. Chapman and Hall, CRC, 1999.
[31] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, no. 5500, pp. 2323-2326, 2000.
[32] J.B. Tenenbaum, V. Silva, and J. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," Science, vol. 290, no. 5500, pp. 2319-2323, 2000.
[33] J.C. Bezdek and R. Hathaway, "VAT: A Tool for Visual Assessment of (Cluster) Tendency," Proc. Int'l Joint Conf. Neural Networks (IJCNN '02), pp. 2225-2230, 2002.
[34] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Proc. Advances in Neural Information Processing Systems (NIPS), 2002.
[35] M. Breitenbach and G. Grudic, "Clustering through Ranking on Manifolds," Proc. 22nd Int'l Conf. Machine Learning (ICML), 2005.
[36] R.B. Cattell, "A Note on Correlation Clusters and Cluster Search Methods," Psychometrika, vol. 9, no. 3, pp. 169-184, 1944.
[37] P. Sneath, "A Computer Approach to Numerical Taxonomy," J. General Microbiology, vol. 17, pp. 201-226, 1957.
[38] T.C. Havens, J.C. Bezdek, J.M. Keller, M. Popescu, and J.M. Huband, "Is VAT Really Single Linkage in Disguise?" Pattern Recognition Letters, 2008, in review.Liang Wang received the PhD

AUTHORS BIOGRAPHY

Asadi Srinivasulu received the B Tech (CSE) from Sri Venkateswara University, Tirupati, India in 2000 and M.Tech with Intelligent Systems in IT from Indian Institute of Information Technology, Allahabad (IIIT) in 2004 and he is pursuing Ph.D in CSE from J.N.T.U.A, Anantapur, India. He has got 10 years of teaching and industrial experience. He served as the Head, Dept of Information Technology, S V College of Engineering, Karakambadi, Tirupati, India during 2007-2009. His areas of interests include Data Mining and Data warehousing, Intelligent Systems, Image Processing, Pattern Recognition, Machine Vision Processing and Cloud Computing. He is a member of IAENG, IACSIT. He has published more than 15 papers in International journals and conferences. Some of his publications appear in IJCA, IJCSET and IJCSIT digital libraries. He visited Malaysia and Singapore.

Dr Ch D V Subba Rao received the B Tech (CSE) from S V University College of Engineering, Tirupati, India in 1991, M.E. (CSE) from M K University, Madurai in 1998 and he was the first Ph.D awardee in CSE from S V University, Tirupati in 2008. He has got 19 years of teaching experience. He served as the Head, Dept of Computer Science and Engineering, S V University College of Engineering, Tirupati, India during 2008-11. His areas of interests include Distributed Systems, Advanced Operating Systems and Advanced Computing. He is a member of IETE, IAENG, CSI and ISTE. He chaired and served as reviewer of IAENG and IASTED international conferences. He has published more than 25 papers in International journals and conferences. Some of his publications appear in IEEE and ACM digital libraries. He visited Austria, Netherlands, Belgium, Hong-Kong, Thailand and Germany.

O.Obulesu received B.Tech degree in Computer Science and Engineering from Sri Venkateswara University in 2005 and M.Tech degree in Computer Science from JNTUA, Anantapur in 2008. He received Gold Medal in M.Tech (Computer Science) Course in the year 2008.He is currently pursuing Ph.D in J.N.T.U.A, Anantapur. He has totally 04 years of experience in Teaching and Industry. Currently working as an Assistant Professor in the Information Technology Department at Sree Vidyanikethan Engineering College (SVNEC), A.Rangampet, Tirupati, Andhrapradesh, INDIA. His research areas are Spatial Data Mining and Spatiotemporal Databases.

P.Sunil Kumar Reddy received MCA from Bharathiar University in 2004 and M.Phil Computer Science from Madurai Kamaraj University. He is pursuing Ph.D in SV University Tirupati. He has 3 years of experience in teaching and 4 years in Industry. Currently working as Senior Software Engineer at Mahindra Satyam, his research areas are Databases and Data Mining.