ISSN:0975-9646

Finding the Number of Clusters in Unlabelled Datasets Using Extended Cluster Count Extraction (ECCE)

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

ABSTRACT: Clustering analysis is the task 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. All clustering algorithms ultimately rely on one or more human inputs, and the most important input is number of clusters (C) to seek. There are many pre and post clustering methods which relieves the user from this choice. These methods ultimately make the choice by thresholding some value in the code. Thus, the choice of c is transferred to the equivalent choice of the hidden threshold that determines C "automatically". 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 n x n image. RDI is generated using VAT (Visual Assessment of Cluster tendency), which highlights potential clusters as a set of "dark blocks" along the diagonal of the image, so that number of clusters can be easily estimated using the number of dark blocks across the diagonal. We develop a new method called "Extended Cluster Count Extraction (ECCE) for counting the number of clusters formed along the diagonal of the RDI.

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

Keywords — Clustering, Cluster Tendency, Reordered Dissimilarity Image, VAT, ECCE, 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 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 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 Cluster Count Extraction (ECCE). 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 ECCE 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.

3. VAT ALGORITHM

The visual assessment of cluster tendency (VAT) tool has been successful in determining potential cluster structure of various data sets, but it can be computationally expensive for large data sets. In this article, we present a new scalable, sample-based version of VAT, which is feasible for large data sets. We include analysis and numerical examples that demonstrate the new scalable VAT algorithm.

3.1. STEPS

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.

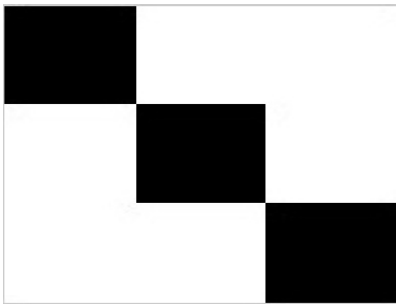

Fig 1 VAT Image

4. ECCE ALGORITHM

The existing system for automatically determining the number of clusters in unlabeled data sets is “cluster count extraction”. The proposed system “Extended Cluster Count Extraction (ECCE)” includes the C-Means algorithm for performing post clustering which improves the accuracy of the existing system.

ECCE algorithm mainly includes five major steps:

- Generating the correlation filter of size s.
- Applying FFT to the VAT image and multiplying with the conjugate filter.
- Applying inverse FFT to the filtered image.
- Compute histogram of off-diagonal pixel values of the back-transformed image.
- Applying C-Means algorithm to the input with number of clusters from above steps.

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.

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.

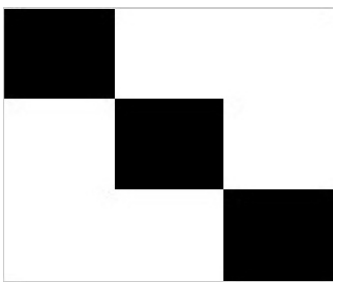

Fig 1(a) VAT Image

Fig 1 (b) FFT Image

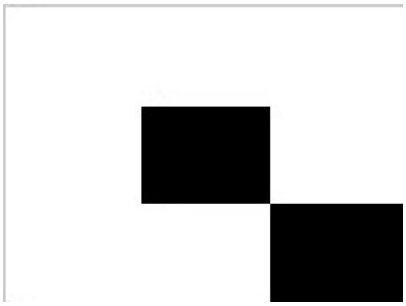

Fig 1(c) FFT Filter Image

When clusters are represented as dark blocks of pixels along the diagonal, choosing a filter with similar characteristics should yield positive results for regions displaying similar characteristics, and negative results for all other regions. To find the largest number of similar characteristics of dark blocks along the diagonal, we apply image segmentation prior to correlation.

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.

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

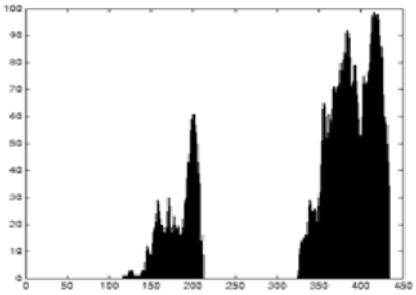

Fig 2 Histogram of image

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.

ECCE ALGORITHM

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

Step 1) Threshold with 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) Put the number of clusters into C-Means Clustering Algorithm and gives very good accuracy.

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

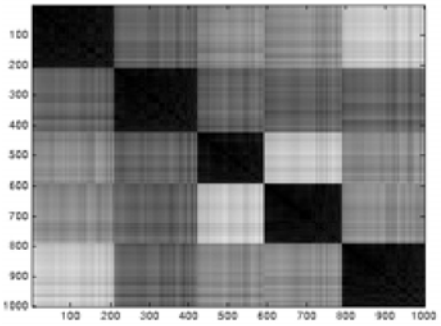

Fig 3 (a)VAT image

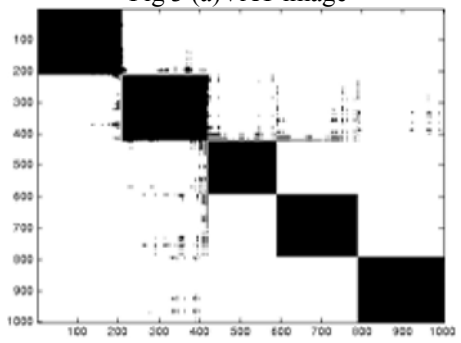

Fig 3 (b) Segmented VAT image

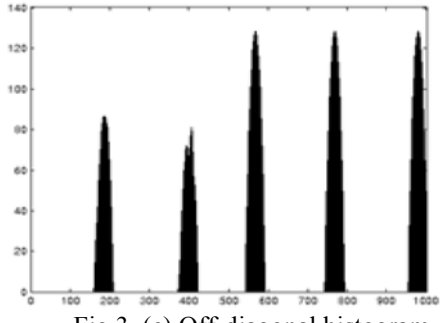

Fig 3 (c) Off-diagonal histogram

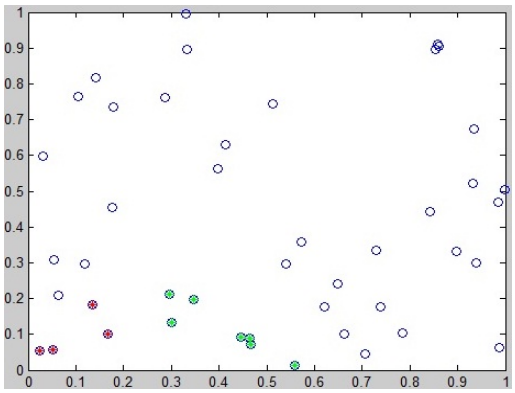

Fig 4 : C-Means algorithm

5. CONCLUSIONS

The proposed technique ECCE replaces the need for human interpretation of the number of clusters in a data set by means of frequency domain correlation to explore the effectiveness of the algorithm. A group of user-generated data sets, and a real-world data example, served as the input to the preliminary cluster detection process. This paper investigates a nearly parameter-free method for automatically estimating the number of clusters in unlabeled data sets. The only user defined parameter is the set S which controls the correlation filter size in ECCE which 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, ECCE compares favorably with post clustering validation methods in computational efficiency. It is noted that ECCE does not eliminate the need for cluster validity, but it simply improves the probability of success. The ECCE technique uses C-Means algorithm to perform post clustering in order to increase the accuracy. A possible extension of this work is to improve the input from data sets to image formats and to perform the clustering on the images. This image clustering can be used in many fields like medical, satellite Imaging and many real world applications.

6. REFERENCES

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

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

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

[9] 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".

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

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

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

[13] 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.

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

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

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

[17] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1998.

[18] J.C. Bezdek, R.J. Hathaway, and J. Huband, "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices," IEEE Trans. Fuzzy Systems, vol. 15, no. 5, pp. 890-903, 2007.

[19] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms," IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

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

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

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 attended, presented and published various National and International conferences. 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, and A.P. 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. Currently working as an Assistant Professor in Information Technology at Sree Vidyanikethan Engineering College, 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.