TPCC: Proxy Based Trusted Pre Cluster Count in Unlabelled Data Sets

Sowmyadevi.k, Janani.R, Mirudula.R, Reha.S, Malini.A, Yatheesha.P

sowmik.me@gmail.com
P.G Scholar, Computer Science Engineering Department,
Coimbatore Institute of engineering and Information Technology, Anna University, India

Abstract- The selection of the number of clusters is an important and challenging issue in cluster analysis. A number of attempts have been made to estimate the number of clusters in a given data set. 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 the number of clusters before clustering occurs. In this paper, we investigate a new method called **Trusted Precluster Count** (**TPCC**) algorithm for automatically estimating the number of clusters in unlabeled data sets, which is based on an existing algorithm **Dark Block Extraction** (**DBE**) of a data set, using several common image and signal processing techniques. Our focus is preclustering tendency assessment. But for completeness we briefly summarize some existing approaches to the post clustering cluster validity problem, before describing visual methods for cluster tendency assessment.

Index Terms- Clustering, cluster tendency, TPCC, DBE

I. INTRODUCTION

Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. It divides a set of objects (data points) into groups (clusters) such that the objects in a cluster are more similar to one another than to the objects in other clusters. The emerging data mining applications place many special requirements on clustering techniques, such as scalability with high dimensionality of data. A number of clustering algorithms have been developed over the decades in data base / data mining community. The selection of the number of clusters is an important and challenging issue in cluster analysis. A number of attempts have been made to estimate the number of clusters in a given data set. 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 the number of clusters before clustering occurs. Our focus is on pre clustering tendency assessment, but for completeness, this project briefly summarizes some existing approaches to the post clustering cluster validity problem, before describing visual methods for cluster tendency assessment.

II. RELATED WORK

Pre-Clustering Tendency Assessment selection of the number of clusters is an important and challenging issue in cluster analysis. A number of attempts have been made to estimate c in a given data set. 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 pre clustering tendency assessment, but for completeness, we briefly summarize some existing approaches to the post clustering cluster validity problem, before describing visual methods for cluster tendency assessment.

To overcome post clustering tendency assessment, Dark Block Extraction (DBE) is introduced for automatically estimating the number of clusters in unlabeled data sets, which is based on the existing algorithm for Visual Assessment of Cluster Tendency (VAT) [1] of a data set, using several common image and signal processing techniques.

III. IMPLEMENTING EXISTING REVAT METHODOLOGY

The Visual Assessment of (cluster) Tendency (VAT) [1] method readily displays cluster tendency for small data sets as grayscale images, but is too computationally costly for larger data sets. A revised version of VAT is presented here that can efficiently be applied to larger collections of data.

Initial Virtual Assessment Tendency (VAT) Algorithm

Input: An $n \times n$ scaled matrix of dissimilarities $D = [d_{ij}]$, with $1 \geq d_{ij} \geq 0; d_{ij} = d_{ji}; d_{ii} = 0$, for $1 \leq i, j \leq n$
Step (i): Set $I = \emptyset$, $J = \{1, 2, \cdots, n\}$ and $P = (0, 0, \cdots, 0)$.
Select $(i, j) \in \arg_{p \in I, q \in J} \max\{d_{pq}\}$.
Set $P(1) = i$, $I \leftarrow \{i\}$ and $J \leftarrow J - \{i\}$.
Step (ii): Repeat for $t = 2, \cdots, n$
Select $(i, j) \in \arg_{p \in I, q \in J} \min\{d_{pq}\}$.
Set $P(t) = j$, update $I \leftarrow I \cup \{j\}$ and $J \leftarrow J - \{j\}$.
Step (iii): Form the re-ordered dissimilarity matrix $\tilde{D} = [\tilde{d}_{ij}] = [d_{P(i)P(j)}]$, for $1 \leq i, j \leq n$.
Output: A scaled gray-scale image $I(\tilde{D})$ so that $\max\{\tilde{d}_{ij}\}$ corresponds to white and $\min\{\tilde{d}_{ij}\}$ to black.

Based on the VAT [1] Algorithm, the reVAT technology is implemented. The purpose of revised VAT (reVAT) is to achieve results similar to VAT with less computation. Revised VAT is implemented with the help of dissimilarity image matrix produced by VAT. Each pixel value is calculated by

using reVAT technology and the pixel result value will be displayed to their corresponding list view column.

IV. IMPLEMENTING BIGVAT METHODOLOGY

Assessment of clustering tendency is an important first step in cluster analysis. One tool for assessing cluster tendency is the Visual Assessment of Tendency (VAT) [1] algorithm. VAT produces an image matrix that can be used for visual assessment of cluster tendency in either relational or object data. However, VAT becomes intractable for large data sets. The revised VAT (reVAT) algorithm reduces the number of computations done by VAT, and replaces the image matrix with a set of profile graphs that are used for the visual assessment step. Thus, reVAT overcomes the large data set problem which encumbers VAT, but presents a new problem: interpretation of the set of reVAT profile graphs becomes very difficult when the number of clusters is large, or there is significant overlap between groups of objects in the data.

BigVAT [2] Algorithm steps include :

1) Choose segmented image from right hand side panel from the segmentation form.

2) Analyze the total number of pixel values and based on that pixel value, create a proxy controller which stores each pixel value of a bigVAT processing image

3)Total Proxy Value Count acts as the threshold value =á‘ from Selected Segmented image ‘M’

4)Considering bigVAT algorithm process, the getpixel( ) function is used to trace pixel value of 0^{th} pixel proxy position and the pixel value will be allowed for dividend into ‘R’, ‘G’, ‘B’ values. Perform a distance transform to obtain a gray scale image and scale the pixel values.

5)Project the pixel values of the image on to the main diagonal axis to form a projection signal. Smooth the signal to obtain the filtered signal by an average filter.

6)And the values will be formulated and the formulated value will be displayed in the corresponding list view position and the calculated pixel is replaced in the same position by using the function called setpixel( ).

7)Once the entire process completed, the output will be displayed in the Screen.

V. IMPLEMENTING DARK BLOCK EXTRACTION METHODOLOGY

One of the major problems in cluster analysis is the determination of the number of clusters in unlabeled data, which is a basic input for most clustering algorithms. In this paper, we investigate a new method called **Dark Block Extraction (DBE)** for automatically estimating the number of clusters in unlabeled data sets, which is based on an existing algorithm for Visual Assessment of Cluster Tendency (VAT) [1] of a data set, using several common image and signal processing techniques.

Dark Block Algorithm steps include:

1) Generate a VAT [1] image of an input dissimilarity matrix,

2) Perform 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.

DBE method is nearly “automatic,” depending on just one easy-to-set parameter.

VI. IMPLEMENTING PROPOSED TRUSTED PRE CLUSTER COUNT METHODOLOGY

TPCC is an advanced method of detecting the number of clusters in a pre defined manner in order to give more accuracy to the segmented image. TPCC is a pre-clustering method, i.e., it does not require the data to be clustered, nor does it find clusters in the data. By using TPCC method, the segmented image is well clearly classified into pixel transformations by maintaining the entire pixel data in a proxy structure ,i.e., in an array format. So the calculations process is very little to find the density of the image in order to produce accuracy to the image. Trusted Pre Cluster Count Algorithm steps include:

1) Generate a VAT image.

2) Perform image segmentation on the VAT image.

3) Divide the image into matrix pixels as row ‘r’ X column ‘c’.

4) Create a proxy controller to store all matrix pixels values.

5) Calculate each pixel value and compare each pixel relate with neighbor pixel.

6) Group the similar result’s pixels .

7) Calculate the number of groups separated which is equal to number of clusters.

The resulted cluster count will be the perfect pre cluster count value where it produces the VAT image into a super quality image.

VII. EXPERIMENTAL RESULTS

Experimental result clearly portraits that implementing Trusted Pre Cluster Count algorithm gives the exact output and the picture quality is improved.

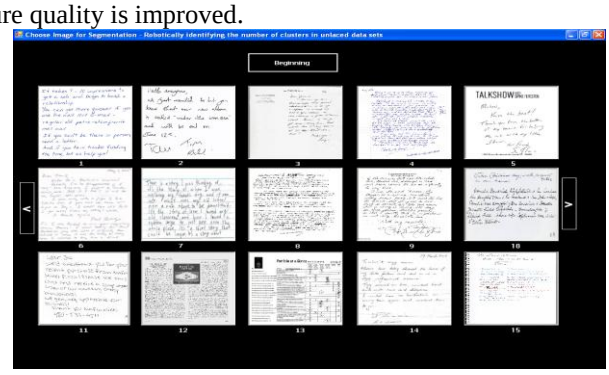

Figure 1: Preview Imaging

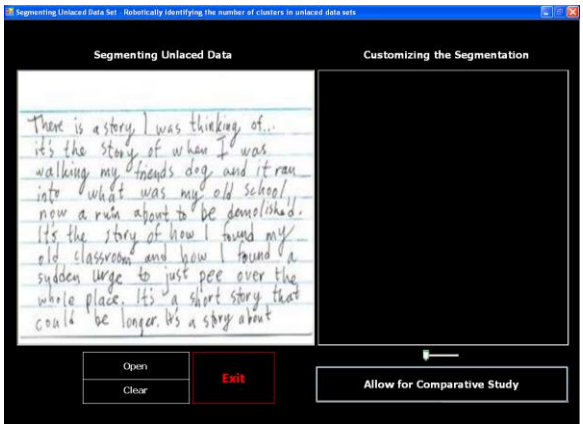

Figure 2: Segmenting Unlaced Data

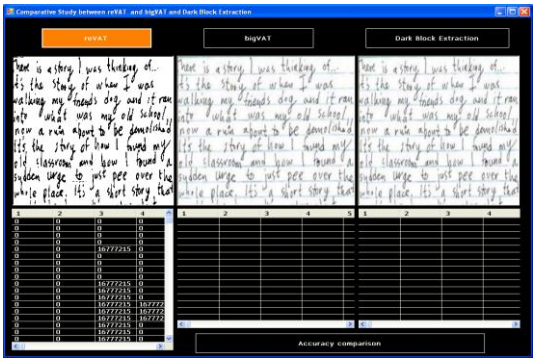

Figure 3: reVAT Output

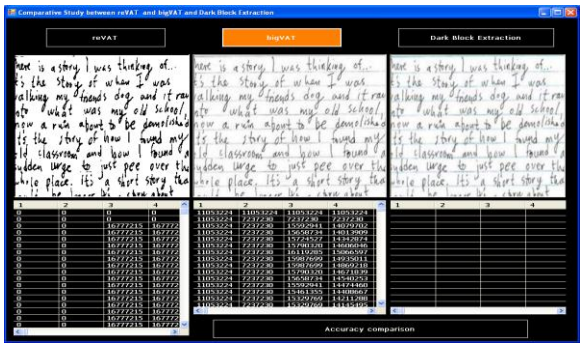

Figure 4: bigVAT Output

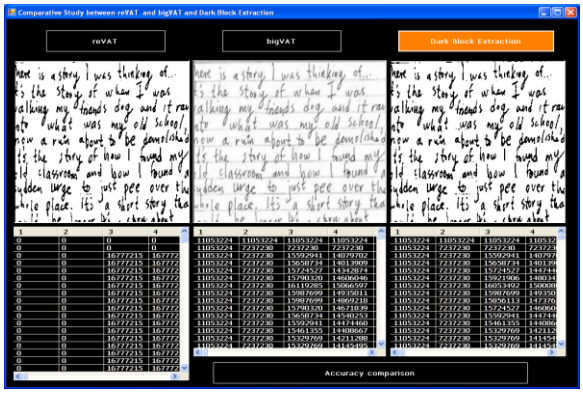

Figure 5 : DBE Output

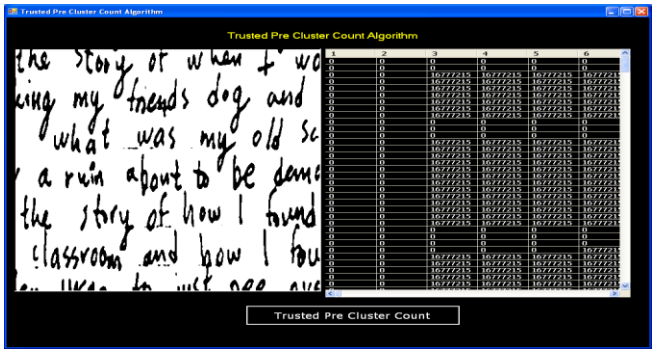

Figure 6 : TPCC Output

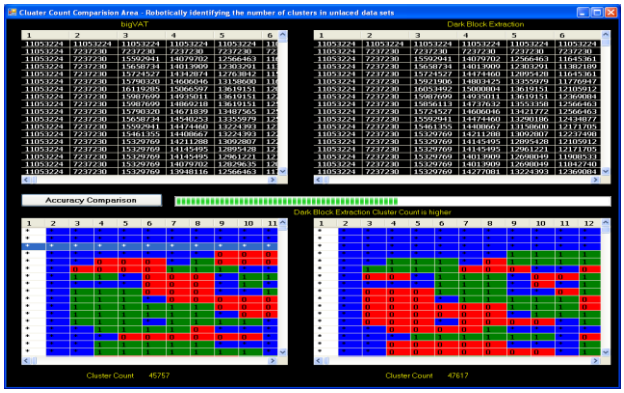

Figure 7 : Cluster Count Comparison of bigVAT and DBE

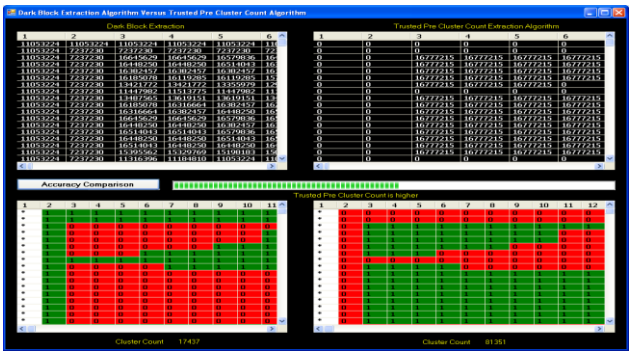

Figure 8: Cluster count Comparison of DBE and TPCC

VIII. CONCLUSION

Experiments confirm what many users of clustering believe: that most methods prefer "larger" rather than "smaller" clusters. Thus the cluster number (corresponding to the detection of major peaks) extracted by DBE appears to be increasingly reliable. This is understandable since if increases, the segmented binary image will be less noisy, which is naturally helpful to subsequent processing. As long as the filter sizes are set to be less than the minimum meaningful cluster size, the larger they are, the more reliable the estimation of the cluster number should be. DBE 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 pairwise dissimilarities when the input data are feature vectors. The euclidean distance may not be suitable for high dimensional or complex data. It is that DBE does not eliminate the need for cluster validity, but it simply improves the probability of success.

IX. FUTURE ENHANCEMENT

DBE is more reliable than CCE. Even though it overcomes the confusing problem in CCE of where to cut the Histogram ,it has Slightly overestimated or under estimated value of 'c', it provides the initial estimation of the cluster number. To overcome this problem , a possible extension of this work concerns the initialization of the **Trusted Pre Cluster Count 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. Extrapolating an approximate centroid for each cluster will not only speed up the termination of the clustering algorithm but also reduce the need for multiple runs with randomized initializations. Inferring the approximate sizes of each cluster. Although DBE is not in and of itself a clustering method, it may provide some useful information on object labels, especially for objects around the peak in the projection signals. If such label information could be used, only the remaining boundary objects need to be clustered, thus reducing the amount of data to be clustered.

REFERENCES

[1] 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.

[2] 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.

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

[4] J.C. Bezdek, R.J. Hathaway, and J. Huband, "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices,"

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

[6] Ping Guo, C. L. Philip Chen, and Michael R. Lyu," Cluster Number Selection for a Small Set of Samples Using the Bayesian Ying–Yang Model" IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 3, 2002.

[7] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James Bezdek, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 3, 2009.

AUTHORS

First Author – Sowmyadevi Kanagasabapathy, Pusuing Master of Engineering in Coimbatore Institute of Engineering and technology, Anna university Coimbatore. As completed B.Tech Information technology in Coimbatore Institute of Engineering and technology, Anna university chennai. Her research work is in the Computer Networking and Data mining.
sowmyaphd@gmail.com

Second Author – **Janani.R** received the Bachelor of Engineering Degree in 2009 and Master of Engineering Degree in 2011.Her research interests are in the area of Networks and Data mining. She is currently working as Assistant professor in Park college, Coimbatore, INDIA
sowmik.me@gmail.com.

Third Author – **Mirudula.R** received the Bachelor of Engineering Degree in 2009 and Master of Engineering Degree in 2011.Her research interests are in the area of Networks and Data mining. She is currently working as Assistant professor in Coimbatore Institute of Engineering and Technology, Coimbatore, INDIA.
sowmyaphd@gmail.com

Fourth Author – **Reha.S** received the Bachelor of Technology Degree in 2010 and pursuing Master of Information Technology Degree 2012. Her research interests are in the area of Networks, Cloud Computing and Data mining. She is currently working as Lecturer in Park College, Coimbatore,INDIA.
rekasubu@gmail.com

Fifth Author – **Malini.A** received the Bachelor of Technology Degree in 2010 and pursuing Master of Information Technology Degree 2012. Her research interests are in the area of Networks, Cloud Computing and Data mining.
swethap.me@gmail.com

Sixth Author – Yatheesha.P received the Bachelor of Engineering Degree in 2010 and pursuing the Master's of Information Technology. Her research interests are in the area of Networks,Network security and Mainframe Networks. She has published a paper on International Journal of Computer Science and Network Security, VOL.12 No.1, January 2012.She is currently working as Lecturer in PPG Insitute Of Technology, Coimbatore,INDIA.
ksowmyabtech@gmail.com.

Correspondence Author – Sowmyadevi.k,
sowmik.me@gmail.com (or) ksowmyabtech@gmail.com.
+91 99428 34269.