Extraction of Dark Blocks to Determine the Number of Clusters in Rectangular Dissimilarity Matrices

Anaswara J Pillai
Dept. of Computer Science
College of Engineering Perumon
Kerala,India

Deepa K Daniel
Assistant Professor
Dept. of Information Technology
College of Engineering Perumon
Kerala,India

Abstract— Cluster tendency Assessment is the problem of determining whether there exist meaningful clusters prior to actual clustering. Cluster analysis is presented to provide the accuracy of clustering objects. This paper proposes a method for extracting dark blocks which determines the number of clusters in rectangular dissimilarity matrix. The proposed algorithm is based on the algorithm co-efficient visual assessment of cluster tendency by using common signal and image processing techniques. First generate an co-efficient iVAT image from a rectangular dissimilarity matrix which is the input of the proposed algorithm. Then the obtained gray scale image is threshold to obtain a binary image. Distance transformation is performed to the binary image to obtain a new gray scale image and the pixel values are projected to the main diagonal axis of image to figure a projection signal. The projected signal is filtered and its first order derivative is computed and the numbers of major peaks obtained in the resulting signal are the number of clusters. Experimental results on Extraction of Dark Blocks in rectangular dissimilarity matrices verify the effectiveness of the proposed scheme.

Index Terms— Dark Block Extraction, Cluster Tendency Assessment, co-clustering, Visualization, Rectangular Dissimilarity Data

I.INTRODUCTION

Clustering is the process of separating the set of objects O={O1,O2,…ON} into self similar groups. There exist a number of clustering algorithms [1].The problem of determining whether there exist meaningful clusters prior to actual clustering is known as cluster tendency assessment. The major problems involved in cluster analysis are:1)Cluster tendency-how many clusters are there(value of c)? 2)Partitioning-grouping the data into c meaningful groups 3)validity-whether the partitions are good?. Many clustering algorithms require number of clusters c as an input parameter. So the quality of clusters largely dependent on the value of c.

Numbers of various techniques for cluster tendency assessment are discussed in [2]. Visual approach for assessing cluster tendency is discussed which is used in all cases involving numerical data. Commonly used one is VAT(Visual Assessment of Cluster Tendency) algorithm [3].All these are about square dissimilarity matrices. Here we are considering rectangular matrices. Assume a rectangular relational data represented by an $m \times n$ matrix D with m row objects O_r and n column objects O_c both combined to form the set O i.e. $N=m+n$ objects .The elements present in D are the pair wise dissimilarities between row and column objects thus known as dissimilarity matrix. Euclidean distance is used as the distance measure. Co-Visual Assessment of Cluster Tendency is a visual approach used for rectangular relational data [4]. There occurs a group of similar objects that consist of only row objects,of only column objects or of mixed objects known as co-clusters. The input of our algorithm is a rectangular dissimilarity matrix D which is the subset of N×N dissimilarity matrix and estimates square matrices D_r, D_c and $D_{r\cup c}$ and their cluster tendency is visually viewed. VAT algorithm is applied on to the D_r, D_c and $D_{r\cup c}$. D is reordered by using $D_{r\cup c}$ and its tendency assessment result in a rectangular co-vat image and that exhibits cluster tendency in D.A path based distance transform in iVAT algorithm is used in co-iVAT algorithm for improving the effectiveness of co-VAT algorithm for showing cluster tendency. But its complexity gets increased. Another algorithm co-efiVAT is implemented which reduces computational complexity in co-iVAT and also improves the ability of co-iVAT to show the cluster tendency in rectangular dissimilarity data.

In this paper a new method is proposed to estimate the number of dark blocks automatically in Reordered dissimilarity images (RDI) of rectangular dissimilarity matrices. Large number of dark block extraction techniques are there.[5][6].The proposed method combines several common image and signal processing techniques[7] and the altered image is projected on the main diagonal axis of RDI. Finally the number of major peaks obtained in the projection signal are the number of clusters.

The rest of this paper is organized as follows: Section II describes the previous work in related domains. Section III presents the proposed algorithm i.e. Extraction of Dark Blocks in co-efiVAT Algorithm. The experimental results are given in Section IV. Finally, a short discussion on conclusions and future study are provide in Section V.

II.RELATED WORK

In this section we discuss existing work related with current work. Visual approaches for various data analysis problems have been widely studied. Visual assessment of cluster tendency algorithm (VAT) is a commonly used one. VAT algorithm is a visual approach for assessing cluster tendency in square dissimilarity matrices. The VAT algorithm shows pair wise dissimilarity between set of objects and displays a reordered form of dissimilarity data. Its reordered dissimilarity image (RDI) is obtained. VAT algorithm is based on prims algorithm for finding the minimum spanning tree of a weighted connected graph[8].The number of dark blocks along the diagonal are the number of clusters.

A major limitation of VAT algorithm is the inability to show cluster tendency when dissimilarity matrix D contains highly complex clusters. It is effective only to show cluster tendency in datasets that contain compact well separated clusters. An improved VAT (iVAT) is proposed[9] which generate RDIs that combine VAT with a path based distance transform. The obtained iVAT images show the number of clusters clearly even if the datasets contain highly complex structure. The distance transform used in the above equation is the shortest path problem solved by using Floyd warshall algorithm[10].

The complexity of the VAT algorithm is $O(N^2)$.The complexity of Floyd warshal algorithm is $O(N^3)$.The total complexity of iVAT is $O(N^3)$.Inorder to reduce this complexity an efficient formulation is proposed in[11].First VAT is applied to the input dissimilarity matrix and then transform the VAT re-ordered dissimilarity matrix into iVAT image using the recursive algorithm efiVAT algorithm.

a) Efficient iVAT Algorithm

```
1. Input:VAT reordered dissimilarity matrix $D^*$ of size $n\times n$.
 2.Parameter:$D'^*=[0]^{n\times n.}$
3.Output:iVAT image.
4.Algorithm:
5.for $r = 2, . . . , n$ do
    6. $j=$ argmin $_{k=1,\ldots,r-1} D^*_{rk}$
    7.$D'^*_{rc=}D^*_{rc}$,c=j
    8. $D'^*_{rc} = \max\{D^*_{rj},D'^*_{jc}\}$,c=1,---r-1,$c\neq j$
6.$D'^*$ is symmetric thus $D'^*_{rc}=D'^*_{cr.}$
7.Obtain iVAT image.
```

Next is the visual cluster tendency assessment of rectangular dissimilarity matrix..The co-VAT algorithm is a visual approach which is used for assessing cluster tendency in m×n rectangular dissimilarity matrix D. D contains dissimilarities between pairs of objects i.e. the first object from O_r and second comes from O_c. O_r is the set of row object i.e. $\{O_1,..,O_m\}$ and O_c the column object set $\{O_{m+1},..,O_{m+n}\}$.The dissimilarity matrix D is of the form:

$$D=\begin{bmatrix} d(O_1,O_{m+1}) & d(O_1,O_{m+2}) & \ldots\ldots & d(O_1,O_{m+n}) \\ d(O2, O_{m+1}) & d(O_2,O_{m+2}) & \ldots\ldots & d(O_2,O_{m+n}) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ d(O2, O_{m+1}) & d(O_2,O_{m+2}) & \ldots\ldots & d(O_2,O_{m+n}) \end{bmatrix} \quad (1)$$

Initially $D_{r\cup c}$ is created in co-VAT algorithm. For creating $D_{r\cup c}$,D_r and D_c is to be estimated. D_r,D_c and $D_{r\cup c}$ are square pairwise dissimilarity matrices for O_r ,O_c and $O_r \cup O_c$ which are the row object set, column object set and the union of row and column object set respectively. For creating $D_{r\cup c}$ first D_r and D_c is to be estimated. Then VAT is applied to D_r,D_c and $D_{r\cup c}$. The matrices $D_{r\cup c}$,D_r and D_c are estimated as given below:

$$D_{r\cup c} = \begin{bmatrix} D_r & D \\ D^T & D_c \end{bmatrix} \quad (2)$$

$$[D_r]_{ij} = \lambda_r\|d_{i*} - d_{j*}\|,\ 1 \le i, j \le m, \quad (3)$$

$$[D_c]_{ij} = \lambda_c\|d_{*\,i} - d_{*j}\|,\ 1 \le i, j \le n, \quad (4)$$

The co-VAT algorithm also has the inability to show the cluster tendency in rectangular dissimilarity data. For that co-iVAT algorithm is used. Before applying VAT algorithm, path based distance transform in (1) is applied to the D_r, D_c and $D_{r\cup c}$. Then VAT algorithm is applied to the transformed square dissimilarity matrixes of D_r, D_c and $D_{r\cup c}$. The problem is that applying the distance transform directly is computationally expensive and its complexity gets increased.

For reducing computational complexity co-efficient iVAT or co-efiVAT algorithm is used. Instead of using distance transform directly a recursive algorithm is used on D_r, D_c and $D_{r\cup c}$[12].First VAT algorithm is applied to the D_r, D_c and $D_{r\cup c}$ and then the recursive algorithm efiVAT is applied to the VAT reordered dissimilarity matrices of D_r, D_c and $D_{r\cup c.}$ Then rectangular co-efiVAT image is built.

Dark block extraction is an algorithm that counts the dark block along the diagonal of an RDI. This method combines several common image and signal processing techniques.First RDIs are generated using VAT algorithm and then image processing operations like segmentation filtering and distance transformation are used. Finally the transformed image is project on to the main diagonal axis of the RDI to form a projection signal i.e. histogram. The number of clusters is estimated by calculating the number of peaks from the histogram. Enhanced dark block extraction Algorithm is also there to count the number of dark blocks using E-VAT algorithm[13].

III.EXTRACTION OF DARK BLOCKS IN RECTANGULAR DISSIMILARITY DATA

A method is proposed to determine the number of dark blocks in rectangular dissimilarity data using co-efficient ivat algorithm(co-effiVAT).This algorithm counts the dark blocks along the diagonal of the RDI. Initially rectangular dissimilarity matrix is taken as the input. Then VAT algorithm is applied on Dr∪c. After that efficient iVAT algorithm is applied on the reordered matrix. The rectangular co-efiVAT image is built. Finally some basic image and signal processing tools are applied on to the formed image.Here thresholding [14], distance transformation, filtering and first order derivative is computed. Its projection signal i.e. a histogram is obtained. The number of major peaks in the histogram is computed as the number of dark blocks.

The main goals of the proposed algorithm are:

- To output the number of clusters very clearly for highly complex data structure in a visual manner.
- To extract the number of dark blocks from a projected signal by counting its number of major peaks.

Some important terms used in the proposed algorithm:

A. *Adaptive Threshold Algorithm.*

Thresholding is the simplest method of image segmentation. It is used to create binary image from a gray scale image. In adaptive threshold algorithm the gray scale image is taken as input and it outputs a binary image representing the segmentation in its simplest execution itself. This technique is successful in tackling the problems of noise and imperfect illumination. Here threshold needs to be computed for every pixel in the image [15].

B. *Distance Transformation.*

A distance transformation is an operation that converts this binary image to a gray-level image where all pixels have a value corresponding to the distance from the pixel to the adjacent non-zero pixel in the binary image [16].A number of DTs are available depending upon which distance metric is being used to determine the distance between pixels. In our proposed algorithm Euclidean distance measure is used. After applying DTs, all pixel values are projected on to the main diagonal axis to form a projection signal histogram which is shown in Fig.8a.

C. *Savitzky-Golay Filtering*

A Savitzky–Golay filter is a digital filter that can be applied to a set of digital data points for the purpose of smoothing the data, that is, to increase the signal-to-noise ratio without greatly distorting the signal[17].Although the histogram is available further smoothing is required to reduce false detections due to noise in the signal. That’s why Savitzky-Golay smoothing filters(digital smoothing polynomial filters or least-squares smoothing filters) is used. The projected signal histogram is shown in Fig.8b.Then first order derivative is computed for better projection of signals and output is shown in Fig.9.

DARK BLOCK EXTRACTION USING CO-EFIVAT ALGORITHM

1. Input: D-m×n rectangular dissimilarity matrix.

2. Output: The number of dark blocks in the RDI i.e., the number of major peaks in the histogram.

3. Build estimates of D_r and D_c using the equations (3) and (4) respectively.

4. Build estimates of $D_{r\cup c}$ using the equation (2).

5. Run VAT on $D_{r\cup c}$ and the reordering indexes are $P_{r\cup c}$= {P(1),…,P(m+n)} and thus $D^{'}_{r\cup c}$ is formed.

6.Run Efficient iVAT Algorithm on $D^{'}_{r\cup c}$ to form $D^{*}_{r\cup c}$.

7.Then its rectangular co-efiVAT image is built from $D^{*}_{r\cup c}$..

8. Reordering indixes R and S are created in such a way that R are the indexes of the elements of $P_{r\cup c}\leq$ m and S the indexes of $P_{r\cup c}$>m.

9. $D^{'*}=[D^{'*}_{ij}]=[(D^{*}_{r\cup c})_{R(i),S(j)}]$,$1\leq i\leq m$ and $1\leq j\leq n$.

10.Repeat the same steps from 5 to 9 for D_r and D_c also to obtain $D^{'*}_{r}$, and $D^{'*}_{c}$,

11. The obtained gray scale image $I_1(D^{'*})$is threshold to obtain binary image I_2’ using adaptive threshold algorithm.

12. Distance transform is performed on the image I_2’ to obtain a new gray scale image I_3’ and the pixel values are projected to the main diagonal axis of image to form a projection signal Histogram H_1.

13. Filter the projected signal H_1 using Savitzky-Golay filter design and new Histogram H_2 is obtained.

14. Compute the first order derivative of H_2 to obtain signal Histogram H_2’.

15.The number of major peaks in the H_2’ is the number of dark blocks i.e. the number of clusters.

The input given is a m×n dissimilarity matrix. The steps 3 to 10 are for applying co-efiVAT algorithm. The obtained RDI is threshold to obtain a binary image and then distance transformation is applied to obtain a new gray scale image. Then these pixel values are projected on to main diagonal axis to form a projection signal histogram. The major peaks in the histogram are calculated as the number of dark blocks. If false detections are present due to noise in the signal further smoothing is required. Here savitzky-golay filtering is used to smooth out a noisy signal. The projection signal histogram finds to be very smooth but requires further smoothing to remove false peaks or detections due to noise in the signal. For that first order derivative of the obtained histogram is calculated and new histogram is obtained. The number of major peaks in the obtained histogram is determined as the number of dark blocks.

IV.RESULTS AND DISCUSSION

In this section, the experiments and the performance results of the proposed algorithm by using one of the UCI data set named IRIS are described. IRIS dataset contains 150 rows and 3 columns. The number of clusters in IRIS dataset is three. Initially the N×N i.e. 150×150 square dissimilarity matrix is calculated. But the input required is m×n rectangular dissimilarity matrix. So the dataset here taken is the subset of IRIS dataset i.e. subset of square dissimilarity matrix calculated. In our numerical example rectangular dissimilarity matrix contains 125 row objects and 148 column objects. The experiment is done using MATLAB. The outcomes of co-VAT, co-iVAT and co-efiVAT image of D'^{*}_{rUc} and its corresponding rectangular images are shown below.co-VAT image of D'^{*}_{rUc} and its rectangular image is shown in Fig.1 and Fig.2 respectively.Fig.3 and Fig.4 shows co-iVAT image of D'^{*}_{rUc} and its rectangular image respectively.Fig.5 and Fig.6 shows the same that belongs to co-efiVAT.

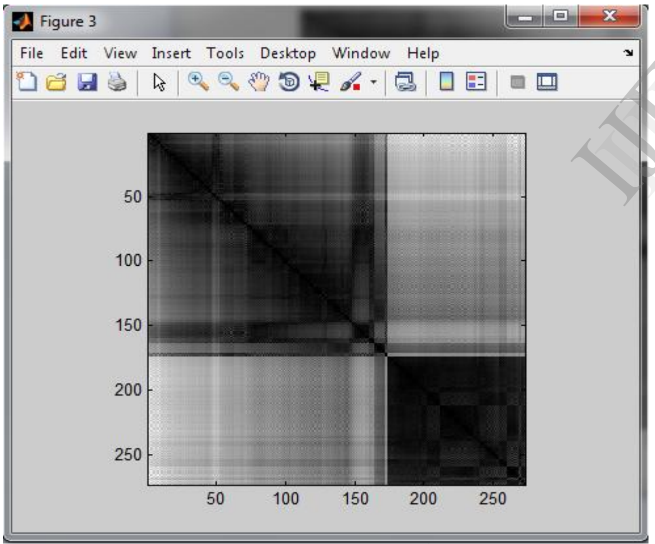

Fig.1.co-VAT image of D'^{*}_{rUc}

From the Fig.1 it is difficult to understand that there exist three clusters visually. But in Fig.2 and Fig.3 it is very clear that there exist three clusters visually.

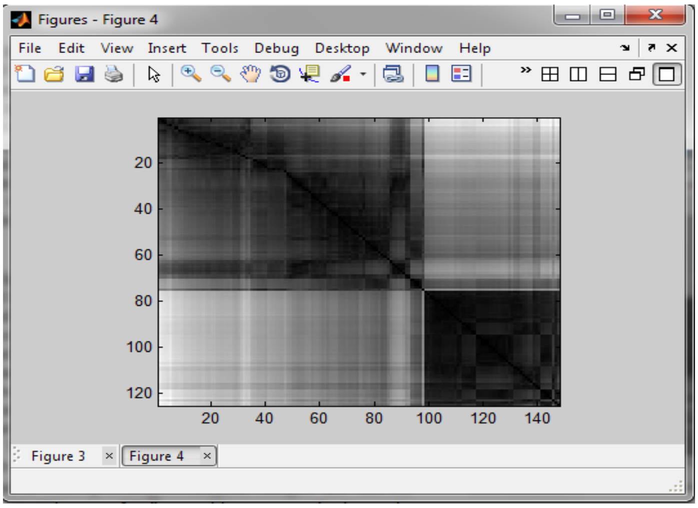

Fig.2.Rectangular co-VAT image

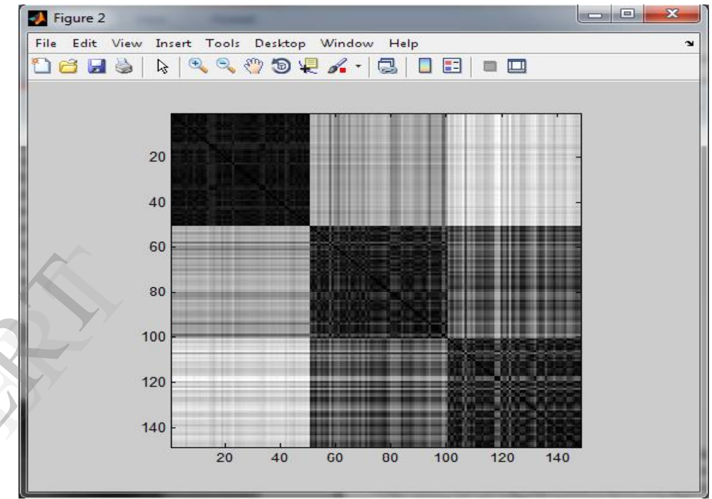

Fig.3.co-iVAT image of D'^{*}_{rUc}

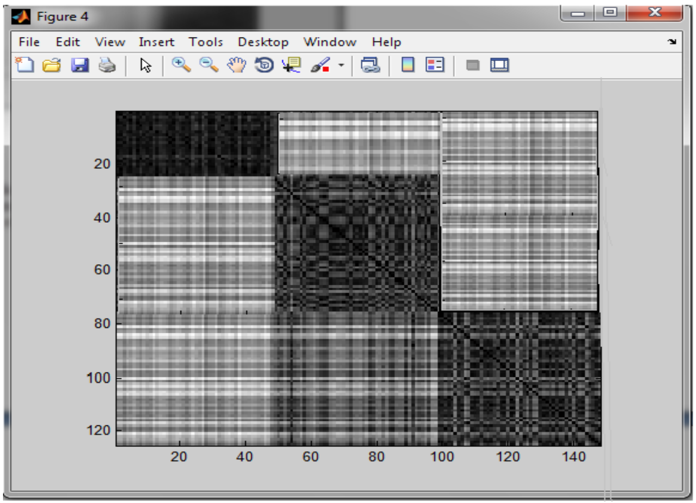

Fig.4.Rectangular co-iVAT image .

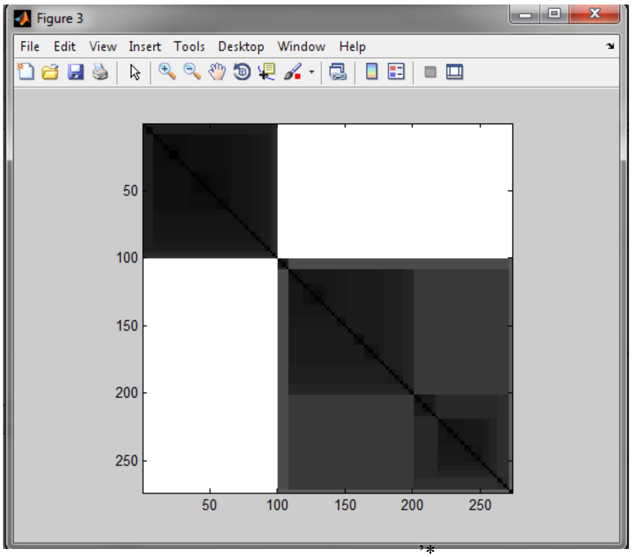

Fig.5.co-efiVAT image of $D'^{*}_{r\cup c}$

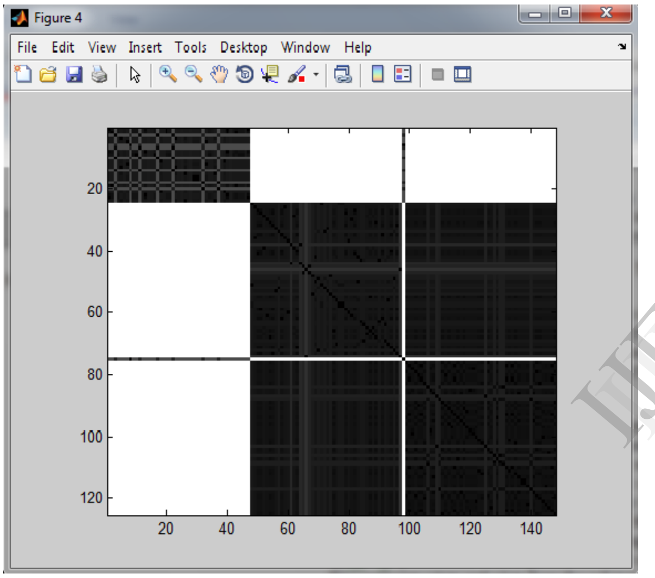

Fig.6 Rectangular.co-efiVAT image

A comparison is done for the run time complexities of co-VAT, co-iVAT and co-efiVAT and it is given in the table below:

	co-VAT	co-iVAT	co-efiVAT
IRIS DATASET	0.688sec	133.631sec	0.712sec

Table1. Run time complexities in co-VAT, co-iVAT and co-efiVAT.

From the above table it is clear that the complexity of co-iVAT algorithm is very high compared to co-VAT and co-efiVAT. The run time complexity of co-iVAT gets increased because of using Floyd warshall algorithm in co-iVAT. The complexity gets reduced by using recursive co-efiVAT algorithm.

In Fig.6 it is easy to determine the number of clusters. The number of dark blocks along the diagonal is the number of clusters that is three. But sometimes it gets confused whether it is three or five.It can be easily determined from our proposed algorithm. By counting the number of major peaks from the projected signal histogram the number of dark blocks is obtained. The outcomes of the steps from 11 to 14 of our proposed algorithm is shown in following figures respectively.

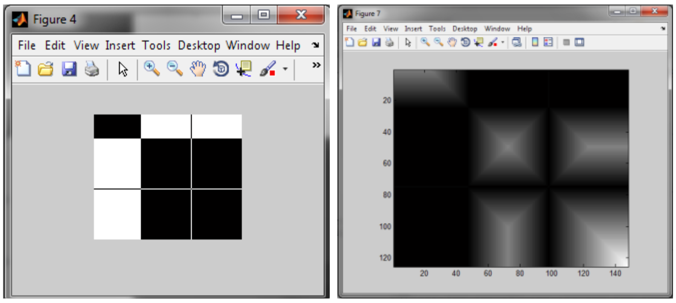

Fig.7.a)Threshold the image in Fig.6 to obtain the binary image I*.b)Distance transformation is applied on I* to obtain new gray scale image I’

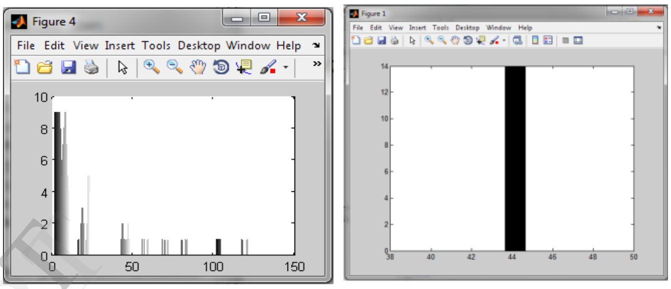

Fig.8.a)Pixel values of I’ is projected to form a histogram H_1.b)Histogram H_2 obtained after filtering the H1 using savitzky- golay filter.

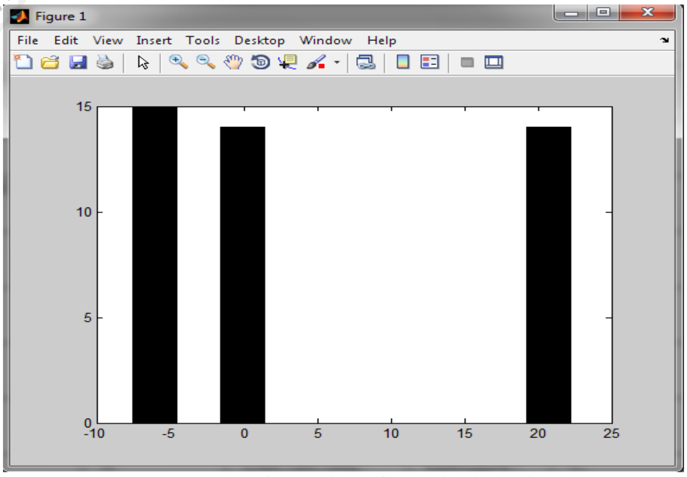

Fig.9.Histogram H_2’ after applying first order derivative to H_2.

Finally the number of clusters is computed as the number of major peaks from the histogram H_2’ in Fig.9.So in our example total number of clusters is three. The proposed algorithm also maintains less run time complexity. The run time complexity obtained for IRIS dataset in our proposed algorithm is 1.02seconds.

V.CONCLUSION

This paper presents a new implementation for extracting dark blocks which determines the number of clusters in rectangular dissimilarity matrix. The existing Dark block

Extraction method works only on square dissimilarity matrix. Here co-efiVAT algorithm is described and dark block extraction is applied on to it.Co-efiVAT algorithm is used to reduce the complexity in co-iVAT algorithm. But the proposed algorithm helps in easier determination of dark blocks from the projected signal histogram. The run time complexities of co-VAT, co-iVAT and co-efiVAT algorithm are compared.

The complexity of the proposed algorithm is also calculated. It is near to the co-efiVAT algorithm but slightly more than that. It can be improved by using some modifications on the proposed algorithm. We are currently examining to replace the computation of $D_{r\cup c}$ with the new alternate reordering scheme. It is computationally less expensive as $D_{r\cup c}$ need not to be constructed. Another extension of this work is to add the initialization of c-means clustering algorithm for object data clustering. This cluster pattern is combined with visual assessment of cluster tendency. By merging these two using an RDI provides a natural environment for visual cluster analysis.

Our Proposed algorithm works only on small datasets i.e. limited to m×n $\approx O(10^4 \times 10^4)$.By extending our proposed algorithm to the scalable versions of co-VAT i.e. scalable co-VAT(sco-VAT) [18] that are too large to be loaded and processed in a single computer which results in an improved visualization of very large data with reduced complexity.

VI. REFERENCES

[1] Rui Xu, Donald Wunsch II, "Survey of clustering algorithms". IEEE Transactions on Neural Networks 16(3), 645–678 (2005).

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

[3] Bezdek, J.C., Hathaway, R.J.: "VAT: A tool for visual assessment of (cluster) tendency" .In: International Joint Conference on Neural Networks, vol. 3, pp. 2225–2230 (2002).

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

[5] L.Wang, C.Leckie K. Rao, and J.Bezdek,"Automatically Determining the Number of Clusters from Unlabeled Data Sets,"IEEE Trans. Knowledge Eng., vol. 21, no. 3, pp. 335-350, Mar. 2009.

[6] Srinivasulu Asadi ,Dr Ch D V Subba Rao ,V Saikrishna "Finding the Number of Clusters in Unlabeled Data sets using Extended Dark Block Extraction," International Journal of Computer Applications (0975 – 8887) Volume 7– No.3, September 2010.

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

[8] K.H. Rosen, "Discrete Mathematics and Its Applications", New York, NY: McGraw-Hill. 1999.

[9] L. Wang, T. Nguyen, J.Bezdek, C.Leckie, and K. Ramamohanarao, "iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment," Proc. PAKDD, June 2010.

[10] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, "Introduction to Algorithms", third ed. MIT Press, 2009.

[11] Timothy C. Havens, James C. Bezdek," An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency(iVAT) Algorithm", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, may 2012.

[12] Timothy C. Havens, James C. Bezdek, and James M. Keller, " A New Implementation of the co-VAT Algorithm for Visual Assessment of Clusters in Rectangular Relational Data", Proc. 10th Int'l Conf. Artificial Intelligence and Soft Computing: Part I (ICAISC '10), pp. 363-371, Apr. 2010.

[13] P. Prabhu, K. Duraiswamy," Enhanced Dark Block Extraction Method Performed Automatically to Determine the Number of Clusters in Unlabeled Data Sets", Int J Comput Commun, ISSN 1841-9836 8(2):275-293, April, 2013.

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

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

[16] Gunilla Borgefors," Distance Transformations in Digital Images", Computer Vision, Graphics, and Image Processing 34, 344-371 (1986).

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

[18] Park, L., Bezdek, J., Leckie, C.: "Visualization of clusters in very large rectangular dissimilarity data", In Gupta, G.S., Mukhopadhyay, S., eds.: Proc. 4th Int. Conf. Autonomous Robots and Agents. (February 2009) 251–256.