ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Clustering of Labeled And Unlabeled Data By Integrating Pre And Post Clustering Approaches

[1] Mr.Asadi Srinivasulu [2] Dr.Ch.D.V.Subbarao [3] I. Muralikrishna
[1]Research Scholar of CSE, [2]Professor of CSE,
[1][2]Department of CSE,
[1]JNTUA, Anantapur,[2][3] S.V.University College of Engineering, Tirupati,

Abstract— Clustering is the process of organizing objects into groups whose members are similar in some way or differ significantly from other objects. There are two approaches viz., pre-clustering and post-clustering. Pre-clustering is an unsupervised learning that assigns labels to objects in unlabeled data. The important pre-clustering approaches that we have considered are Dark Block Extraction (DBE), Cluster Count Extraction (CCE) and Co-VAT (Visual Assessment of Cluster Tendency). The present work focuses on pre-clustering approach. The limitations of these pre-clustering algorithms are i) DBE can't handle the large data ii) CCE suffers because of perplexing iii) Co-VAT works with only rectangular data. Our work proposes Extended Dark Block Extraction (EDBE), Extended Cluster Count Extraction (ECCE) and Extended co-VAT to overcome the above said limitations. The following five steps results after integrating pre and post clustering approaches. They are 1) Extracting 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 smoothing the pixel values on the main diagonal axis of the image to form a smoothening signal. 4) Applying first-order derivative and fast fourier transformation on smoothened signal for detecting major peaks and valleys. 5) Now post-clustering approach i.e. k-means algorithm is applied to the major peaks and valleys in-order to obtain refined clusters. The proposed algorithms viz., EDBE, ECCE and Extended Co-VAT uses VAT as well as the combination of several image processing techniques are applied on various real world data sets like IRIS, WINE and Image Data sets. These extended approaches use Reordered Dissimilarity Image (RDI) that highlights potential clusters as a set of 'Dark blocks' along the diagonal of the image. The simulation results show that EDBE, ECCE, Extended co-VAT outperform DBE, CCE and co-VAT in terms of time-complexity and accuracy of labeled and unlabeled data.

Keywords: **Clustering, DBE, CCE,CO-VAT, VAT, iVAT, EDBE, ECCE and Extended CO-VAT.**

I. INTRODUCTION

1.1 Introduction to Pre-clustering

Pre-clustering tendency assessment is a process of finding the number of clusters in data sets, which is an important and challenging issue. Pre-clustering is an approach suggested by Huse et al. 2010. A common problem in the data mining community is how to organize the observed data into meaningful structures. As an exploratory data analysis tool, cluster analysis aims at forming objects of similar kind into their respective groups. Several clustering algorithms have been studied and are mentioned in the literature survey. In general, clustering of unlabeled data pose many problems like assessing cluster tendency, i.e., how many clusters to form or what is the value of 'cluster count', partitioning the data into clusters, validating the cluster count and cluster performance i.e. how to increase the quality. Several attempts have been made to estimate number of clusters in a given data set. All these methods are used to identify the validity of the clusters, i.e., they try to select the best partition among all the alternatives. In contrast, tendency assessment attempts to estimate 'cluster count' before clustering occurs.

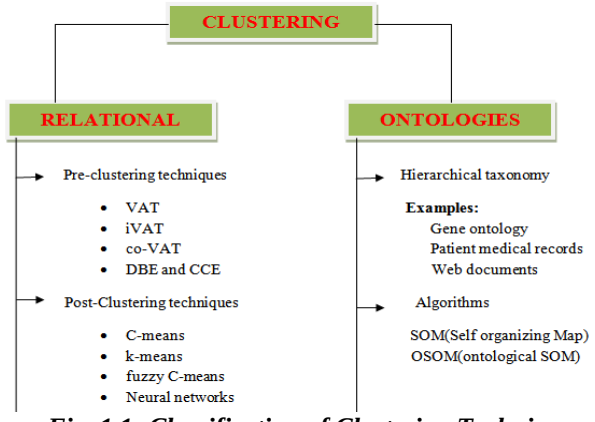

Fig. 1.1: Classification of Clustering Techniques

There are large numbers of clustering algorithms reported in the literature such as VAT, iVAT, DBE, CCE,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Co-VAT. In general, clustering of labeled and unlabeled data pose four major problems: 1) Assessing cluster tendency, i.e., how many clusters to ask for or what is the 'c' value 2) Partitioning the data into meaningful groups 3) Validating the discovered clusters 'c' and 4) Increasing the quality of the clusters. The VAT algorithm addresses this problem by reordering the dissimilarity matrix D so that, the number of clusters is identified by just viewing the diagonal axis of an image, iVAT improves the contrast of the VAT images. The results obtained are useful to easily identify the cluster tendency. Rectangular data is the special form of relational data and it is the dissimilarity value between a set of row objects Or and a set of column objects Oc, where the relation among row or column objects is unknown. Co-VAT algorithm is a new formulation and it is a visual method for identifying the cluster tendency of the rectangular data.

1.2 Visual Assessment of Cluster Tendency (VAT)

The VAT algorithm addresses the problem i.e. poor cluster performance on images by reordering the dissimilarity matrix D so that, the number of clusters is displayed as the number of "Dark Blocks" along the diagonal. The reordered dissimilarity image will highlight the potential clusters by just observing the number of dark blocks that are present in the diagonal axis of an image. The RDI image will provide the number of clusters by just identifying the dark blocks that are present in the diagonal axis of an image; each cluster is different from the other. This dissimilarity matrix generated is provided as input to the VAT algorithm. The Reordered Dissimilarity Image provides a potential cluster structure from the pair wise dissimilarity matrix of the data by using VAT. Region segmentation, directional morphological filtering, and distance transformation are the image processing operations used to segment the regions of interest in the RDI to convert the filtered image into a distance transformed image. The number of clusters can be easily extracted from the one dimensional signal which is nothing but the diagonal axis of the reordered dissimilarity image. These extracted clusters use sequential signal processing operations like average smoothing and peak detection. These operations are used to found the peaks and valleys from the work signal and those are made to satisfy certain conditions. The curves which satisfy the given conditions will only be considered as the valid points.

1.3 Dark Block Extraction (DBE)

In DBE, simple euclidean space is used to calculate pairwise dissimilarities when the input records are feature vectors. The euclidean distance may not be appropriate for high dimensional or composite data. The results of DBE are not clear and the cluster extraction is not performed accurately as DBE is based on VAT. DBE cannot handle the large datasets. In short, DBE counts the dark blocks along the diagonal of an RDI. By integrating DBE with k-means algorithm, it is modified as Extended Dark Block Extraction (EDBE). Hence EDBE assures a good quality dissimilarity measures for diverse types of given large data sets.

1.4 Cluster Count Extraction (CCE)

In CCE, the histogram is generated by first thresholding and then applying the 2-Dimensional Fast Fourier Transformation (FFT). It is further continued by window correlation in the frequency domain, later by back-transforming to the spatial domain and finally by performing the correlated primer off-diagonal histogram. The positions of peaks and valleys in DBE implicitly correspond to centers and ranges of sub blocks (or clusters). It is hard to observe similar phenomenon from the CCE histograms. CCE also counts dark blocks in RDIs using a combination of several image processing techniques. K-Means algorithm increases the accuracy of the CCE compared to that of DBE.

1.5 Co-VAT

In Co-VAT, it takes only rectangular dissimilarity matrix of size m × n matrix D, where the elements of D are pair-wise dissimilarities between m row objects Or and n column objects Oc. The union (Dr ∪ Dc) of these disjoint sets of (N = m + n) objects O. Clustering tendency assessment is a process by which a dataset is analyzed to determine the number(s) of clusters present. Co-Visual Assessment of Tendency (Co-VAT) algorithm is proposed for rectangular data. Co-VAT first inputs pair-wise dissimilar values among the row objects, and renders a square relational matrix Dr and Dc for the column objects and then builds a larger square dissimilarity matrix Dr ∪ Dc. The clustering questions can be addressed by using the VAT algorithm on Dr, Dc, and Dr ∪ Dc, where D is reordered by shuffling the reordering indices of Dr ∪ Dc

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

II. LITERATURE SURVEY

In this chapter the existing pre-clustering addressed in the literature are discussed briefly. Clustering approaches are broadly classified into two categories, which are pre-clustering and post-clustering. The process of grouping data into subgroup classifications prior to clustering has been described as 'pre-clustering'. Post-Clustering is "the process of organizing objects into groups whose members are similar in some way". Each approach is explained elaborately. The criterion of user, supplying the number of clusters to be formed increases the complexity and also it is highly expensive, because the same clustering algorithm needs to be implemented many times with different user inputs. There by, it follows some pre-clustering approaches like CCE, DBE and co-VAT [3] to estimate the number of clusters before clustering. In this chapter, the literature review of various pre-clustering approaches has been discussed. In the existing studies the number of pre-clustering is not considered in prior to the clustering process. However, a completely satisfactory solution (best number of clusters and comparative study) is not available in the literature.

Bezdek & Hathaway (2002) implemented a VAT algorithm [8] for visually identifying the number of clusters. It is used only a pair wise dissimilarity matrix D is needed as the input. When the vectorial forms of objects are presented, it is easy to convert them into D using dissimilarity methods. Even when vectorial data are not clearly presented, it is still possible to use certain metrics to calculate a pair wise dissimilarity matrix. The VAT [8] image shows the number of clusters and estimated members of object, matrix reordering produces in a hierarchy of clusters.

Liang Wang et al (2009) described a new Dark Block Extraction (DBE) method for identifying the number of clusters in large amount of datasets automatically. It is used in many common image and signal handing techniques with the existing VAT algorithm [8]. First the dissimilarity matrix is formed from the VAT [8] image, next apply the segmentation processing in the image. Next step is to apply a distance transforming method to an image and representing the pixel values to diagonal axis of the image. Finally we get number of clusters based on the smoothing projection signal.

Liang Wang et al (2010) defined improved visual representation for finding number of clusters automatically. It handles the complex datasets and hidden objects using spectral analysis of the final clusters. First it converts the given dataset into pair wise dissimilarity matrix. From the matrix, we form an image of each object that is reordered as dark blocks along the diagonal to reveal hidden cluster structure. The algorithm unsuccessful in highly complex datasets which are not possible to highlight the cluster structure and also visual data partitioning, index based cluster validation is not supported. This method clearly shows the cluster results which depend on the selection of input criteria.

Havens & Bezdek (2012) defined graph-theoretic distance transform method called improved VAT (iVAT) algorithm [2], to overcome the issues of VAT algorithm [8]. This method is a combination of VAT image and iVAT [2] image which are extensively reduces the computational complexity. This method finds the VAT reordered dissimilarity matrix, in this the distance transform functions is applied. The iVAT [2] algorithm is to display very large dissimilarity images due to resolution limitations which are imposed by current graphics. In order to solve all the issues of above mentioned approaches, a new extended approach is introduced for automatically identifying the number of clusters in numerical, character and image data.

III. MOTIVATION AND OBJECTIVES

One of the major problems in cluster analysis is to determine the number of clusters in unlabeled data prior to clustering. The major limitations of these pre-clustering algorithms are: i) DBE can't handle the large data ii) CCE suffers from perplexing iii) Co-VAT works with only rectangular data. The above limitations are the key motivations for extending pre-clustering approaches and they are as described: DBE can't handle the large data: DBE with VAT algorithm effectively handles smaller data sets of size 50x50 or less than that. So DBE works only with smaller matrices. This remains as a limitation with DBE. CCE suffers from perplexing: In CCE, after correlation filter is computed to clean the VAT image which is done by multiplying the complex conjugate of the filter to the VAT image by applying the Fast Fourier

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Transform. After performing image processing techniques, CCE generates histograms. These histograms contain valleys and peaks, where to cut the histogram valleys and peaks i.e. perplexing problem.

Co-VAT works with only rectangular data: Co-VAT first computes row objects pair-wise dissimilarity values (i.e. the rectangular relational matrix Dr) and the column objects (i.e. the rectangular relational matrix Dc), and then builds a larger dissimilarity matrix (i.e. the rectangular relational matrix Druc). Here it is observed that all the above said matrices are rectangular relational matrices. So Co-VAT works only with rectangular matrices. This is the limitation of Co-VAT.

To solve the above said drawbacks, new extended approaches are proposed. These proposed approaches overcome the shortcomings mentioned above and also produces better result compared to existing approaches.
Objectives of the proposed work

- To find the number of clusters in IRIS, WINE and IMAGE data sets into homogeneous and distinct groups.
- To identify the group of similar objects and to discover division of patterns and correlations in large data sets.
- To integrate pre and post-clustering approaches to improve the quality and performance of a cluster count.
- To estimate the number of clusters in both the labelled and unlabeled data prior to the application of clustering mechanism.

The salient contributions of the research work are:

- The Extended Dark Block Extraction (EDBE) is proposed to overcome the limitation of DBE, as DBE can't handle the large data.
- The Extended Cluster Count Extraction (ECCE) is proposed to overcome the limitation of CCE which suffers from perplexing.
- The Extended co-VAT is proposed to overcome the limitation of Co-VAT which works only with rectangular data.

IV. PROPOSED APPROACHES

Diverse pre-clustering algorithms have been proposed in this work for efficient clustering process, i.e., EDBE, ECCE and Extended co-VAT. Clustering objects can be represented by means of labeled and unlabeled data. The proposed approaches in this work perform efficient clustering. The clustering process may result in different partitioning of a data set, depending on the specific criterion used for clustering. Thus there is a need of preprocessing the dataset before we initiate the clustering process. The existing system determines the number of clusters in unlabeled data sets with limitations like perplexing, and its inability in overlapping of histograms is overcome by the proposed techniques. The proposed approaches are developed to automatically determine the number of clusters in labeled and unlabeled datasets.

4.1 Extended Dark Block Extraction (EDBE)

EDBE aims to overcome the limitations of DBE. In DBE, simple euclidean space is used to calculate pairwise dissimilarities when the input records are feature vectors. The euclidean distance may not be appropriate for high dimensional or composite data. The results of DBE are not clear and the cluster extraction is not performed accurately as DBE is based on VAT. DBE cannot handle the large datasets. In short, DBE counts the dark blocks along the diagonal of an RDI. EDBE results by integrating DBE with k-means algorithm. Hence EDBE assures a good quality dissimilarity measures for diverse types of given large data sets.

The limitations of DBE is summarized as follows

- DBE cannot handle larger datasets.
- DBE uses complex techniques for smoothing and filtering.
- DBE does not overcome the problem of perplexing in CCE.
- DBE is less reliable than CCE.
- DBE yields modest better accuracy than CCE.
- EDBE is proposed to overcome the above said drawbacks of DBE.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

The basic steps involved in EDBE are as listed below

1. Dissimilarity Transformation and Image segmentation: As the information about possible cluster in the data is embodied in the dark blocks in the RDI, an important preprocessing mechanism is used to extract the regions of interest.
2. Directional Morphological filtering of binary image: To make the segmented image clearer, especially for the cases in which the degree of overlap between-clusters is large, morphological operations is used to perform binary image filtering.
3. Distance transformation and diagonal smoothening of filtered image: In order to convert the morphologically filtered image into an informative one that clearly shows the dark block structure information, it is necessary to consider the values of pixels that are along or off the main diagonal axis of the image.
4. Detection of major peaks and valleys in the smoothed signal: The number of dark blocks in any RDI is equivalent to the number of "major peaks" in the smoothening signal.
5. Applying the k-means algorithm to the major peaks: The k-means algorithm is applied to major peaks and valleys to generate accurate number of clusters.

The detailed algorithm of EDBE is as given below

EDBE Algorithm

1) Start
2) Find the threshold value 'α' from 'm' using Otsu's algorithm.
3) Transform 'm' in to new dissimilarity matrix 'm1' with m1ij = 1- exp(-m/α).
4) Form an RDI image 'I1' using the previous module.
5) Threshold 'I1' to obtain a binary image 'I2' using algorithm of Otsu.
6) Filter 'I2'using morphological operations to obtain a filtered binary image 'I3'.
7) Perform a distance transform on 'I3' to obtain a gray scale image 'I4'and scale the pixel values to [0, 1].
8) Project the pixel values of the image on to the main diagonal axis of 'I4' to form a projection signal 'H1'.
9) Smoothen the signal 'H1' to obtain the filtered signal 'H2' by an average filter.
10) Compute the first order derivative of 'H2' to obtain 'H3'.
11) Find peak position 'pi' and valley positions 'vj' in 'H3'.
12) Select valid peaks by considering some conditions and number of valid peaks which gives number of clusters.
13) Put the number of clusters into k-means clustering algorithm and gives very good accuracy.
14) Stop

Performance Evaluation of EDBE

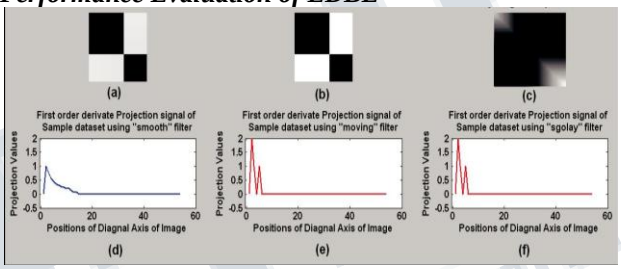

Fig. 4.1(a),(b),(c),(d),(e),(f): EDBE with k-means Algorithm

The above Fig. 4.1(a) shows the number of dark blocks after computing first order derivative projection signal of sample dataset using "smoothing" filter. Fig. 4.1(b) shows the number of dark blocks after computing first order derivative projection signal of sample dataset using "moving" filter. And in Fig. 4.1(c) it is evident that the number of dark blocks after computing first order derivative projection signal of sample dataset is calculated using "sgolay" filter. Fig.4.1 (d)(e)(f) shows the peaks in the graphs between positions of Diagonal axis of image and Projection value.

Merits of EDBE

- EDBE handles the larger data sets.
- EDBE overcomes the problem of perplexing in CCE.
- EDBE is independent of the representation of data to be clustered, so long as a pair wise distance matrix is available to represent the data.
- EDBE handles the problem of contiguous overlapping regions through morphological operations.
- EDBE uses simple techniques for smoothing and filtering.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

4.2 Extended Cluster Count Extraction (ECCE)

ECCE aims to overcome the limitations of CCE. In CCE, the histogram is generated by first thresholding and then applying the 2-Dimensional Fast Fourier Transformation (FFT). It is further continued by window correlation in the frequency domain, later by back-transforming to the spatial domain and finally by performing the correlated primer off-diagonal histogram. The positions of peaks and valleys in DBE implicitly correspond to centers and ranges of sub blocks (or clusters). It is hard to observe similar phenomenon from the CCE histograms. CCE also counts dark blocks in RDIs using a combination of several image processing techniques. K-Means algorithm increases the accuracy of the CCE compared to that of DBE.

The limitations of CCE is summarized as follows

- CCE faces the problem of perplexing (where to cut the histogram).
- CCE is much less reliable than DBE.
- The results of CCE are less accurate than DBE.

The detailed algorithm of ECCE is as given below

ECCE Algorithm

1) Start
2) Threshold the RDI image with Otsu's algorithm.
3) Choose a correlation filter ratio of size s.
4) Apply the Fast Fourier Transform (FFT) to both the segmented RDI and the filter.
5) Multiply the transformed RDI with the complex conjugate of the transformed filter.
6) Compute the inverse FFT for the filtered image.
7) Take the off-diagonal pixel values (e.g., qth off diagonal) of the back-transformed image and compute its histogram.
8) Cut the histogram at an arbitrary horizontal line y ¼ v and count the number of spikes.
9) Put the number of clusters into k-Means Clustering Algorithm as it gives very good accuracy.
10) Stop

Performance Evaluation of ECCE

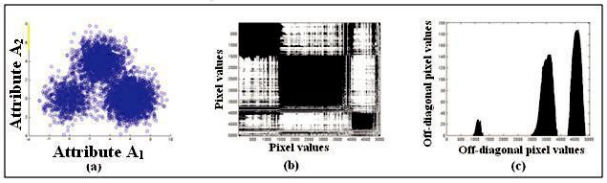

Fig. 4.2(a),(b),(c): Results of ECCE approach (a) Scatter plot of sample data set (b) Ordered VAT image (c) ECCE histogram

The above Fig. 4.2(a) displays the number of clusters over attributes A1 with attribute A2 after computing Fast Fourier Transformation (FFT) projection signal of sample dataset using "smoothing" filter. The above Fig. 4.2(b) shows the number of dark blocks over pixel values of attribute A1 and A2. After computing Fast Fourier Transformation (FFT) projection signal of sample dataset using "moving" filter. The above Fig.4.2(c) displays the number of valleys over off-diagonal pixel values after computing Fast Fourier Transformation (FFT) projection signal of sample dataset using "smoothing" filter.

Merits of ECCE

- ECCE overcomes the problem of perplexing (where to cut the histogram).
- ECCE is much more reliable than CCE.
- The results of ECCE are more accurate than CCE.
- k-Means algorithm increases the accuracy of the ECCE to that of CCE

4.3 Extended Co-VAT

Extended Co-VAT aims to overcome the limitations of Co-VAT. In Co-VAT, it takes only rectangular dissimilarity matrix D of size m × n, where the elements of D are pair-wise dissimilarities between m row objects Or and n column objects Oc. The union of these disjoint sets Dr ∪ Dc consists of N objects where N = m + n. Clustering tendency assessment is a process by which a dataset is analyzed to determine the number(s) of clusters present. Co-Visual Assessment of Tendency (Co-VAT) algorithm is proposed for rectangular data. Co-VAT is a visual approach that addresses four clustering tendency questions:

i) How many clusters are in the row objects?
ii) How many clusters are in the column objects Oc?
iii) How many clusters are in the union of the row and column objects Or ∪ Oc?
iv) How many (co)-clusters are there that contain at least one of each type?

Co-VAT first inputs pair-wise dissimilar values among the row objects, and renders a square relational matrix Dr and Dc for the column objects and then builds a larger square dissimilarity matrix Dr ∪ Dc. The clustering questions can be addressed by using the VAT algorithm on Dr, Dc, and Dr ∪ Dc, where D is reordered by shuffling the reordering indices of Dr ∪ Dc.

The limitations of co-VAT is summarized as follows

- Co-VAT works only on rectangular datasets.
- Computation time is more.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

- Shows nominal cluster tendency of images in "tough" cases (i.e. does not comply with large image datasets).

The detailed algorithm of Extended Co-VAT is as given below

Extended co-VAT Algorithm

1) Start
2) Input the rectangular matrix D (m*n).
3) Build estimates of Dr and Dc.
4) Build estimate of Dr ∪ Dc.
5) Run VAT on Dr ∪ Dc.
6) Form the Co-VAT ordered Rectangular dissimilarity matrix D*.
7) Forms a Reordered dissimilarity matrices images of D*, D*r , D*c ,and D*r∪Dc
8) Stop

Performance Evaluation of Extended Co-VAT

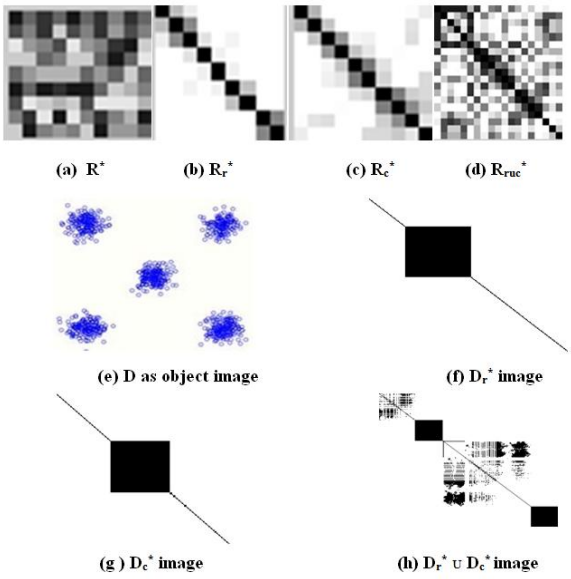

Fig. 4.4(a),(b),(c),(d),(e),(f),(g),(h): Results of applying Extended Co-VAT algorithm on image dataset

The above Fig 4.4(a)(b)(c)(d)(e) show the number of clusters over 'D' (Dissimilarity) as object image after computing dissimilarity matrix. It is also observed that Fig 4.4(f) is obtained from Dissimilarity row wise computation. Fig 4.4(g) is obtained from dissimilarity column wise computation and Fig 4.4(h) is obtained from dissimilarity row union column matrix computation.

Merits of Extended co-VAT

- Extended Co-VAT works both for square and rectangular matrices m x n (or) n x n
- Fast computation time.
- Shows good cluster tendency of images in "tough" cases (i.e. does comply with large image datasets).
- It works for any size of datasets.

5. RESULTS OF PROPOSED APPROACHES

In order to evaluate the performance of proposed approaches, we consider various types of data sets such as numerical, character and image.

5.1 Numerical datasets and Image Data sets:

The following data sets are taken for comparative study

1. IRIS Data
2. WINE Data
3. IMAGE Data

5.1.1 IRIS dataset:

Iris dataset may consist of five attributes such as SL, SW, PL, PW and the Class, where Class represents three types such as IRIS Setosa, Iris Versicolor, and IRIS Virginica. The sepal length, sepal width, petal length, and petal width is determined in centimeters. This dataset may consist of 150 instances where each class will have 50 instances. This IRIS dataset was obtained from the UCI website, available on this link, https://archive.ics.uci.edu/ml/datasets/Iris. This dataset has no missing values. The following may represent the summary statistics of the IRIS data.

5.1.2 WINE data:

The attributes are filtered pair wise and dissimilarity matrix is computed by using the visual assessment of cluster tendency algorithm and also the Euclidean distance is calculated. The results obtained by applying the DBE to this wine data, where the number of clusters formed is equal to 3. The first order derivatives at the values of α as 0.01 and 0.03, which obtains the c value as 3. The smoother appearance of the signal is obtained by the larger filter values.

5.1.3 IMAGE Data:

This face dataset was first used in the paper clustering through ranking on manifolds, which is extracted from the Yale-B face dataset. This data consists of only single light source image. The images are in different poses and with different conditions. The total of 576 views of an image is present in this dataset. The background details related to every subject in the 9 poses are captured. The number of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

images (n) that are present in this dataset is 1755. Each and every image may consist of 30×40 number of pixels.

5.2 Extended Dark Block Extraction (EDBE)

The EDBE automatically estimates the number of clusters in unlabeled data sets. The dark block extraction method combines both the images and the signal processing techniques. The reordered dissimilarity matrix is obtained by using VAT algorithm which is computed by using the Euclidean distances. It shows that the data contains only two geometric clusters. By applying this IRIS dataset to the EDBE algorithm, it is known that the number of clusters obtained is equal to three. This is an encouraging result from both the geometric or physical point of view. The EDBE is one of the methods which is used to estimate the number of clusters in an unlabeled data. This approach estimates the number of clusters by diagonally counting the number of dark blocks in a reordered dissimilarity image.

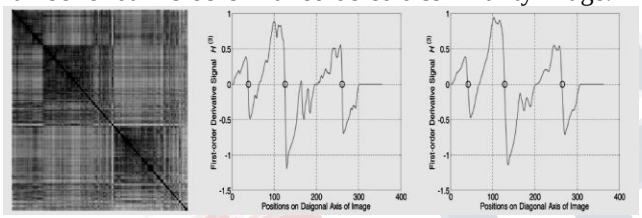

Fig. (a) Fig. (b)

Fig. 5.1(a), (b): Results on the wine data

In the above mentioned figures, the two main dark blocks indicate that the value of cluster count i.e. 'c' is 2. By observing the diagonal of an image, it is clear that dark blocks are related to the geometric clusters. The larger dark block in this VAT image consists of two subspecies. The above Fig. 5.1(a) indicates that the number of clusters that are formed is equal to 3 as where Fig. 5.1(b) indicates the number of clusters as two.

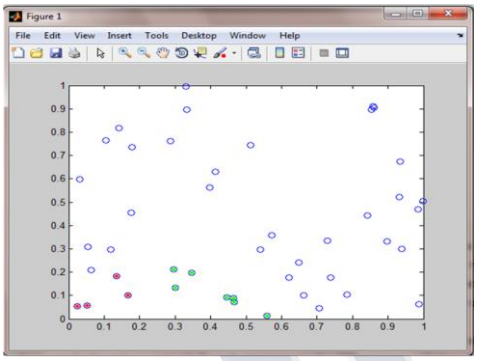

Fig 5.2: Graph for EDBE with k-Means Algorithm

Fig 5.2 displays the number of clusters as three after integrating pre clustering approach EDBE with k-means algorithm. The colors of these clusters are indicated in green, red and blue.

5.3 Extended Cluster Count Extraction (ECCE)

The Extended Cluster Count Extraction is used to count the number of clusters in unlabelled datasets which uses the VAT algorithm and combination of some image processing techniques. This algorithm also counts the number of dark block diagonally in an image by using FFT (Fast Fourier Transform) which is one of the methods that are used in this algorithm.

The major steps involved here is to compute the FFT and also the inverse of FFT.

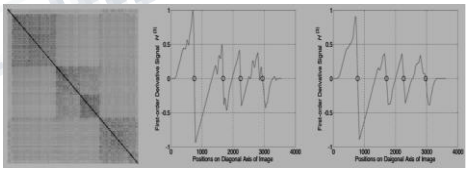

Fig 5.3: Results on image data

The results in Fig 5.3 are obtained by applying the visual assessment of cluster tendency algorithm to the Face dataset. We can observe that in the VAT image, the middle dark block in the diagonal axis is divided into two small blocks. By applying the DBE algorithm to this data, the number of clusters obtained is equal to four, even though the number of groups that are present in this data is equal to three.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

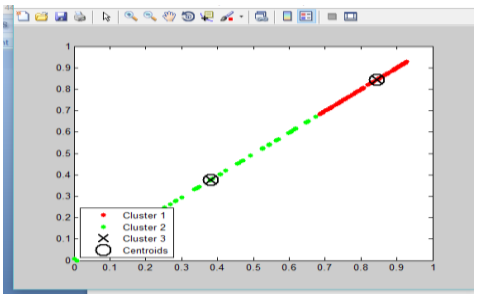

Fig 5.4: Graph for ECCE with k-means Algorithm

Fig 5.4 displays the number of clusters as three after integrating pre clustering approach ECCE with k-means algorithm. The colors of these clusters are indicated in green, red and blue.

5.4 Extended Co-VAT

The Extended Co-VAT algorithm displays the cluster tendency of the multiple types of rectangular data clusters before submitting the k-means Algorithm. It is the Co-visual assessment of tendency that is mainly proposed for the rectangular data.

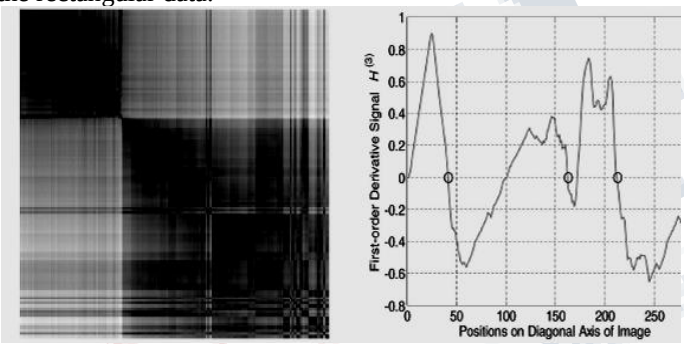

Fig 5.5: Results on the *IRIS* data

Co-VAT first computes row objects pair-wise dissimilarity values (i.e. the rectangular relational matrix D_r) and the column objects (i.e. the rectangular relational matrix D_c), and then builds a larger dissimilarity matrix (i.e. the rectangular relational matrix $D_r \cup D_c$). After computing the above three steps of data, Fig 5.5 displays the dark block on diagonal, after computing first-order derivative signal over the dark blocks. The number of clusters displayed is shown on diagonal axis of image.

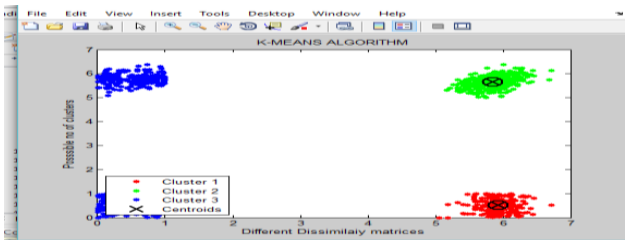

Fig 5.6: Graph for Extended Co-VAT k-means Algorithm

Fig. 5.6 displays the number of clusters as three after integrating pre clustering approach i.e. Co-VAT with k-means algorithm. The colors of these clusters are shown in green, red and blue.

6. COMPARATIVE STUDY

This section discusses the comparative study of the proposed approaches viz., Extended Dark Block Extraction (EDBE), Extended Cluster Count Extraction (ECCE), & Extended Co-VAT approaches.

Dataset	Physical Classes	Attributes	Size (n)	Number of Clusters C		
				Extended Co-VAT	ECCE	EDBE
N_1	2	3	3*3	3	3	3
N_2	3	10	10*10	3	3	3
N_3	3	50	5000* 5000	3	3	3
N_4	5	100	100000* 100000	3	3	3
I_1	2	3	3*3	3	3	3
I_2	3	10	10*10	3	3	3
I_3	3	50	5000* 5000	3	3	3
I_4	5	100	100000* 100000	3	3	3

Table 6.1: Data sets used and the number of clusters *C* for the proposed approaches

where N_1 – N_4 stands for numerical data sets, I_1- I_4 is image data sets.

The data set characteristics viz., size of datasets, physical classes, attributes, number of clusters of EDBE, ECCE and Extended Co-VAT are summarized in Table 6.1.

Evaluation Metrics: The performance of the proposed approaches is calculated using constraint partitioning k-means algorithm and they are computed by means of three evaluation measures.

1. Clustering Accuracy (C_a).
2. Clustering Error (C_e).
3. Time Complexity (O(n))

The above evaluation metrics are used in the proposed approaches

6.1 Clustering Accuracy: Clustering accuracy is used for estimating the performance, which is given as follows**:**

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

$c_a = \frac{1}{N_d} \sum_{i=1}^{N_c} N_{cc}$

where, N_d = Number of data points in the dataset.
N_c = Number of resultant clusters.
N_{cc}= Number of data points occurring in both cluster 'N' and its
corresponding class.

In-order to determine the cluster tendency of data using proposed approaches, it can group the data based on characteristics using VAT. Correlation of incidence and VAT image results for the proposed approaches viz., EDBE,ECCE and Extended Co-VAT with integration of k-means algorithm for two data sets is as given below.

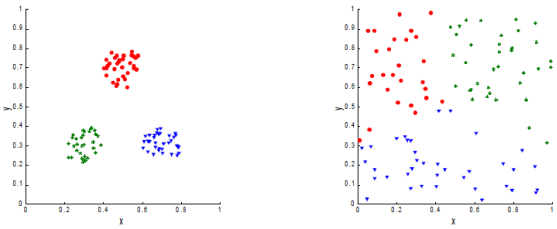

Fig. 6.1: Number of clusters of *IRIS* and *WINE* data sets

Fig 6.1 displays three clusters after integrating pre-clustering approaches viz., DBE, CCE and Co-VAT with k-means algorithm. The colors of the three clusters are indicated in green, red and blue.

5.2 Clustering Error: Cluster 1 is the same as Cluster 2, but comparing Cluster 1 and Cluster 2 element-by-element leads to Clustering Error.
Clustering Error is $C_e = 1 - C_a$
where, C_e = Clustering Error .
C_a = Clustering Accuracy.

5.3 Time Complexity: The code corresponding to EDBE,ECCE and Extended Co-VAT is run on the given data set and it will take 1.130691 sec, 2.130691 sec and 3.130691 sec respectively to obtain the number of clusters. The following Table 6.2 discusses the comparative study of the proposed approaches with existing approaches.

Pre-Clustering Approach	#Instances	#Attributes	#Clusters	Cluster Accuracy	Cluster Error(α)	Time Complexity
DBE	150	4	2	72.35	0.15	2.1306 Sec
CCE	150	4	2	60.10	1.25	3.1306 Sec
Co-VAT	150	4	2	65.00	1.50	4.1306 Sec
EDBE	500	8	3	92.35	0.15	1.1306 Sec
ECCE	500	8	3	80.10	1.25	2.1306 Sec
Extended co-VAT	500	8	3	85.00	1.50	3.1306 Sec

Table 6.2: Comparison of Real Data Sets characteristics and results using Existing vs. Proposed Approaches

In Table 6.2, several real-time examples are considered to evaluate the performance of proposed approaches viz., EDBE, ECCE, and Extended Co-VAT. The tested three examples viz., *IRIS*, *WINE* and *IMAGE* data is from UCI Machine learning repository. In each data set, VAT image is calculated and the corresponding first-order derivative and Fast Fourier Transformation is also computed. The first order derivative signal is calculated with α=0.03 value for comparison of different 'α' values in Cluster Error(α). The proposed approaches show results with parameters such as cluster accuracy, cluster error and computational time. The extended pre-clustering approaches are compared with existing approaches by taking the above said parameters into account and which significantly reduced the computational complexity of estimation of cluster count when compared with the existing DBE, CCE, Co-VAT algorithms. The integration of pre and post-clustering techniques results in better computational cost and cluster count accuracy.

7. CONCLUSION AND FUTURE WORK

The novel algorithms that have been represented in this thesis estimates the number of clusters in labeled and unlabeled datasets. The proposed pre-clustering techniques i.e., Extended Dark Block Extraction (EDBE), Extended Cluster Count Extraction (ECCE), Extended Co-VAT approaches overcomes the limitations of DBE, CCE and Co-VAT. The limitations include i) Dark Block Extraction (DBE) can't handle the large data, ii) Cluster Count Extraction (CCE) suffers from perplexing, iii) co-VAT works with only rectangular data.
The extended pre-clustering algorithms improved the quality of VAT images, which ideally increased the

interpretability of clustering tendency. The post-clustering approach i.e. k-means resulted into accurate output of number of clusters. The comparative study of existing pre-clustering approaches viz., VAT, Co-VAT, DBE, CCE with extended approaches viz., Extended Co-VAT, EDBE and ECCE has been carried out by simulation and taken various sized numerical and real time image datasets (*IRIS* and *WINE*) as input. The outcome of comparative study is that EDBE shows better performance and computational efficiency than ECCE and Extended Co-VAT. These extended pre-clustering techniques are compared with existing approaches by taking the following parameters into account i) Cluster Error ii) Cluster Accuracy and iii) Time Complexity and which significantly reduced the computational complexity of estimation of cluster count when compared with the existing DBE, CCE, Co-VAT algorithms. The integration of pre and post-clustering techniques results into better computational cost and cluster count accuracy and have produced a large amount of new knowledge resources which leads to new research directions for big data sets such as Facebook and Twitter.

REFERENCES

[1] Liang Wang, Christopher Leckie, Kotagiri Ramamohanarao, and James Bezdek, "Automatically Determining the Number of Clusters in Unlabeled Data Sets", Vol. 21, No. 3, pp. 335-350, Fellow, IEEE- March 2009 .

[2] Timothy C. Havens, Senior Member, IEEE, and James C. Bezdek, "An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm", Vol. 21, No. 3, pp. 335-350, Fellow, IEEE, 2012.

[3] Timothy C. Havens1, James C. Bezdek1, and James M. Keller1, "A New Implementation of the co-VAT Algorithm for Visual Assessment of Clusters in Rectangular Relational Data", Vol. 21, No. 3, pp. 335-350, Fellow, IEEE, 2012.

[4] Ahmad A, Dey L (2007) K-Mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering63: 503-527, Volume 63, Issue 2, November 2007.

[5] Azuaje F, Dubitzky W, Black N, Adamson K (2000) Discovering relevance knowledge in data: a growing cell structures approach. IEEE transactions on systems, man, and cybernetics Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society30: 448-460, Volume: 30 Issue: 3, Jun 2000.

[6] Bandyopadhyay S, Saha S (2008) A point symmetry-based clustering technique for automatic evolution of clusters. Knowledge and Data Engineering, IEEE Transactions on20: 1441-1457, Volume: 20, Issue: 11, Nov. 2008.

[7] Belkin M, Niyogi P (2001) Laplacian Eigen maps and spectral techniques for embedding and clustering. Advances in neural information processing systems14: 585-591, Volume 15 Issue 6, June 2003.

[8] Bezdek JC, Hathaway RJ (2002) VAT: A tool for visual assessment of (cluster) tendency. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on, Vol. 3, pp 2225-2230, 2002.

[9] Bezdek JC, Hathaway RJ, Huband JM (2007) Visual assessment of clustering tendency for rectangular dissimilarity matrices. Fuzzy Systems, IEEE Transactions on15: 890-903, Volume: 15 Issue: 5, Oct. 2007.

[10] Bezdek JC, Pal NR (1998), some new indexes of cluster validity. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on28: 301-315, Volume: 28 Issue: 3, Jun 1998.

[11] Breitenbach M, Grudic GZ (2005) Clustering through ranking on manifolds. In Proceedings of the 22nd international conference on Machine learning, pp 73-80, doi:10.1016/j.neucom.2009.03.012, & 2009.

[12] Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Communications in Statistics-theory and Methods3: 1-27, Volume: 41, Issue: 12, 2012.

[13] Cattell R (1944) A note on correlation clusters and cluster search methods. Psychometrika 9: 169-184, Volume 9, Issue 3, September 1944.

[14] Cross VV, Sudkamp TA (2002) Similarity and compatibility in fuzzy set theory: Assessment and Applications, Vol. 93: Physical Verlag, 2002.

[15] Czekanowski J (1909) Zur differential diagnose derNeand ertalgruppe: Friedr. Vieg & Sohn, DOI: 10.1371/journal.pone.0136550, September 29, 2015.

[16] Dhillon IS, Modha DS, Spangler WS (1998) Visualizing class structure of multidimensional

data. Computing Science and Statistics: 488-493, volume = "30", year = "1998".

[17] Floodgate G, Hayes P (1963) The Adansonian taxonomy of some yellow pigmented marine bacteria. Journal of General Microbiology30: 237-244, Volume 30, Issue 2, 1963.

[18] Garai G, Chaudhuri B (2004) A novel genetic algorithm for automatic clustering. Pattern Recognition Letters25: 173-187, Volume 25 Issue 2, 19 January 2004.

[19] Girolami M (2002) Mercer kernel-based clustering in feature space. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council13: 780-784, Volume: 13 Issue: 3, May 2002.

[20] Gonzalez RC, Woods RE, Eddins SL (2009) Digital image processing using MATLAB, Vol. 2: Gatesmark Publishing Tennessee, 2009.

[21] Grünwald P, Kontkanen P, Myllymaki P, Silander T, Tirri H (1998) Minimum encoding approaches for predictive modeling. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pp 183-192, 1998.