VISUALIZATION OF CLUSTERS USING scoivat ALGORITHM

Malarvizhi.M,
Department of CSE,
Srinivasan Engineering College,
Perambalur,India.
crabmalar@yahoo.co.in

Jayanthi S,
Assistant professor
Department of CSE
Srinivasan engineering college,
Perambalur,India
nigilakash@gmail.com

Abstract- **Assessment of cluster tendency is an important step in cluster analysis. Clustering is the problem of partitioning a set of unlabeled objects O = {o1,…, on} into self-similar groups. In conventional (object data) cluster analysis, the objects are separated into groups according to their features. Although clustering is typically thought of as only the act of separating objects into the proper groups, cluster analysis actually consists of three concise questions: Cluster tendency, Partitioning and Cluster validity. One tool for assessing cluster tendency is the Visual assessment of cluster tendency(VAT) algorithm. The VAT algorithm is a visual method for determining the possible number of clusters in, or the cluster tendency of a set of objects.VAT produces an image matrix that can be used for visual assessment cluster tendency in either relational or object data. An efficient formulation of the iVAT algorithm is used which significantly reduces its computational complexity. iVAT implementation begins by finding the VAT reordered dissimilarity matrix and then performs a distance transform on this matrix. The iVAT is performed only on the small datasets. To improve the clustering accuracy and effectiveness of the iVAT scoiVAT algorithm is used. scoiVAT algorithm improves the cluster tendency and it supports large datasets. scoiVAT clustering algorithm can be used in various applications such as malignant tumors, genes expressed in a microarray experiment.**

Key words- clustering, cluster tendency, partitioning, iVAT, scoiVAT.

I. INTRODUCTION

Data mining is used to extract the hidden predictive information from large databases. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining is used in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace. Data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources. Data mining automates the process of finding predictive information in large databases. Data mining tools sweep through databases and identify previously hidden patterns.

II. PROBLEM STATEMENT

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. The appropriate clustering algorithm and parameter depend on the individual data set and intended use of the results.

Determining clusters have been one of the most popular platforms for solving data mining problems. However, building and managing clusters exhibits several drawbacks: 1) Fixed number of clusters can make it difficult to predict; 2) Does not work well with unlabelled data sets; and 3) Different initial partitions can result in different final clusters.

Regarding these limitations, data mining technology has been proposed as a viable solution to determine clusters, to improve the clustering accuracy and tendency. In a recent work, the clustering algorithm extends this by including visualization of clusters, so providing a flexible and agile management of clustering.

The VAT algorithm is a visual method for determining the possible number of clusters in, or the cluster tendency of a set of objects. The improved VAT (iVAT) algorithm uses a graph-theoretic distance transform to improve the effectiveness of the VAT algorithm for “tough” cases where VAT fails to accurately show the cluster tendency. Existing system presented an efficient formulation of the iVAT algorithm which reduces the computational complexity of the iVAT algorithm from O(N3) to O(N2). iVAT used Euclidean distance for dissimilarity matrix computation. It also proves a direct relationship between the VAT image and the iVAT image produced by our efficient formulation.

The major drawbacks are 1) Computing the symmetric matrix was confused. Since many clustering stratergy contains complex and inefficient algorithms. 2) Poor cluster tendency. Dissimilarity matrix calculation is complex which reduces the cluster tendency. 3) Only used for small datasets. Scalability is the major requirement which is lacking in the existing system.

III. RELATED WORK

A. *Visual assessment of cluster tendency(VAT)*

A method is given for visually assessing the cluster tendency of a set of Objects 0 = {ol, ..., on} when they are represented either as object vectors or by numerical pair wise dissimilarity values. The objects are reordered and the reordered matrix of pair wise object dissimilarities is displayed as an intensity image. Clusters are indicated by dark blocks of pixels along the diagonal. The problem of determining whether clusters are present as a step prior to actual clustering is called the assessing of clustering tendency. The VAT approach presents pair wise dissimilarity information about the set of objects 0 = {o, ,.. ., on} as a square digital image with pixels. After the objects are suitably reordered the image is better able to highlight potential cluster structure. 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 0[2].

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 0. 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 has missing components (is incomplete), then any existing data imputation scheme can be used to “fill in” the missing part of the data prior to processing. The ultimate purpose of imputing data here is simply to get a very rough picture of the cluster tendency in 0. For incomplete object data a pair wise Euclidean (or other norm) dissimilarity R can be used, from incomplete xi and xj simply by using all features common to both object data, and then properly scaling the result, based on how many of the s possible features are actually used.

The following are several points about VAT:

Only a pair wise dissimilarity matrix is required as the input for the VAT algorithm. When vectorial forms of object data are available, it is easy to convert them into a dissimilarity matrix using any vector norm. Even when vectorial data are unavailable (e.g., when clustering sequences of different lengths), it is still feasible to use some flexible dissimilarity metrics to convert them into pair wise relational data, e.g., using Dynamic Time Warping (DTW) for sequence matching for measuring the dissimilarity between two point sets of different sizes. VAT is used to estimate the number of clusters prior to clustering. Even if the estimated result does not coincide with the true (but unknown) value, it provides a basis for setting a suitable range into which the correct number of clusters may fall. VAT depends only on the input D, so a “good” D is critical when D is derived from object vectors. If the input data have high dimensionality and/or is nonlinearly separable, it may be better to calculate D in a compact feature space after (nonlinear) feature extraction, rather than in the original input space[4].

B. *Self-organizing visual assessment of cluster tendency(SO-VAT)*

Cluster analysis or clustering is the assignment of a set of data samples into subsets (called clusters) in such a form that data in the same cluster are similar in some sense. One of the major problems in cluster analysis is the determination of the number of clusters in unlabelled data. Spectral-VAT algorithm has been used combining spectral analysis and to automatically determinate the number of cluster, this involves the eigen-decomposition of an nxn similarity matrix, which is clearly intractable for a large number (n) of samples .The VAT algorithm is based on the principle that cluster structure in an unlabeled data set may be revealed by an image of some reordering of the rows and columns of the dissimilarity matrix, resulting blocks in the ordered image that correspond to clusters in the data[3].

The Self-Organizing Map (SOM) is a type of artificial neural network trained by unsupervised learning to produce a low dimensional discrete representation of the training data distribution, called a map, usually configured as a two dimensional grid of neurons. It is a powerful tool in data mining, as it is capable of projecting high-dimensional data onto a neuron grid with good topological preservation between both spaces. This method present a new algorithm, SO-VAT (Self-Organizing Visual Assessment of cluster

Tendency), to deal with large data. The new algorithm models the data using a SOM, then selects a group of the neuron prototypes according to their density of activation, and then applies the VAT algorithm to the selected prototypes. This algorithm takes advantage of the use of the SOM, reducing significantly the computation time when applying the VAT algorithm. An important characteristic of the SOM is that input-space regions with a number of data samples are represented with a proportional number of neurons, thus high-density data regions have more detailed prototype-representation than sparse-data regions. Therefore, removing neurons with low density of activation is a good way to get a cleaner prototype-representation for clustering[3].

The simplest way to define a density matrix on the SOM is the Hit Histogram, which visualizes density by counting how many input samples are assigned to each neuron prototype. This solution lacks of taking into account the prototype inter-distances. A more accurate representation of the density of activation is the P-Matrix. As calculating an exact Voronoi volume is a difficult task, they use approximated volumes of hyper-spheres of certain radius, whose centers are the neuron prototypes. To simplify the calculation it follow a different approach: each neuron prototype and its adjacent prototypes in the map grid, as if they were positioned locally in a plane in the input space. After ordering the map neurons by their densities, it proceed to the elimination of the less relevant neurons: those neurons with low density-values are removed from the map.

C. *Clustering in ordered dissimilarity data(CLODD)*

It presents a new technique for clustering either object or relational data. First, the data are represented as a matrix D of dissimilarity values. D is reordered to D□ using a visual assessment of cluster tendency algorithm. If the data contain clusters, they are suggested by visually apparent dark squares arrayed along the main diagonal of an image I (D□) of D□. The suggested clusters in the object set underlying the reordered relational data are found by defining an objective function that recognizes this blocky structure in the reordered data. The objective function is optimized when the boundaries in I (D□) are matched by those in an aligned partition of the objects. The objective function combines measures of contrast and edginess and is optimized by particle swarm optimization.

This prove that the set of aligned partitions is exponentially smaller than the set of partitions that needs to be searched if clusters are sought in D. Clustering in unlabeled data X or D is the assignment of labels to the objects in O that are groups of similar items. The two necessary ingredients of all attempts to cluster in X or D are the number of groups to seek and (a model that encapsulates) some mathematical way to assess or assign similarity between the various objects.

CLODD is a completely autonomous method for determining cluster tendency, extracting clusters from the image of the reordered dissimilarity data, and providing a cluster validity metric, as well. This leads to a distinct advantage of CLODD, that CLODD is not tied directly to any one distance metric or reordering scheme. CLODD requires, as input, only an image of reordered dissimilarity data, such that the clusters appear as dark blocks along the diagonal[5].

D. *Extended dark block extraction(EDBE)*

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. This method focus 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 EDBE is introduced, it mainly includes two algorithms, i.e. Visual Assessment of Cluster Tendency (VAT) and Extended Dark Block Extraction (EDBE). Here, it initially concentrates 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, EDBE is applied[1] .

E. *Visual assessment of cluster tendency using diagonal tracing(VATDT)*

In- stead of displaying the ordered dissimilarity matrix (ODM) as a 2D gray-level image for human interpretation as is done by VAT, it trace the changes in dissimilarities along the diagonal of the ODM. This changes the 2D data structure (matrices) into 1D arrays, displayed as a tendency curves, which enables one to concentrate only on one variable, namely the height. One of these curves, called the d-curve, clearly shows the existence of cluster structure as patterns in peaks and valleys, which can be caught not only by human eyes but also by the computer. Numerical experiments showed that the computer can catch cluster structures from the d-curve even in some cases where the human eyes see no structure from the visual outputs of VAT[1].

VATdt algorithm is meant to replace the straightforward visual displaying part of the VAT algorithms. It can start from an ordered dissimilarity matrix from any algorithm of that kind. Instead of displaying the matrix as a 2-dimensional gray-level image ODI for hu- man interpretation, VATdt analyzes the matrix by taking averages of various kinds along its diagonal and produces the tendency curves, with the most useful of them being the d-curve. This changes 2D data (a matrix) into a 1D array, which is certainly easier to both human eyes and the computer since the concentration is now only on one variable the height. Possible cluster structure is reflected as high-low pat- terns on the d-curve with a relatively uniform range that enables the computer to catch them with thresholds[7].

IV. SYSTEM ARCHITECTURE

This is the general architecture that represents the high performance clustering using the proposed algorithms.

To improve the clustering accuracy and effectiveness of the VAT, scoiVAT algorithm is applied. It present an efficient formulation of the scoiVAT algorithm which reduces the computational complexity. It have an m*n matrix D, and that its entries correspond to pair wise dissimilarities between m row objects Or and n column objects Oc, which, taken together (as a union), comprise a set O of N= m+n objects. It develops a new visual approach that applies to four different cluster assessment problems associated with O.

The scoiVAT is the combination of scalableVAT, coVAT and improvedVAT algorithms. The scalableVAT is used to improve the scalability of the datasets. It allows very large 2D object datasets as an input that contains x and y coordinates. These coordinates are plotted into an image that contains 2D image. Then dissimilarity matrix is calculated using mahalanobis algorithm. This distance transform algorithm calculates the dissimilarity values between the coordinate points.

The coVAT algorithm is then applied to the dissimilarity matrix. The row objects are first clustered using coVAT clustering algorithm in the following way

$$[D_r]_{ij} = |d_i^* - d_j^*|, \text{ for } 1 \leq i \leq m, 1 \leq j \leq m;$$

The row object clustering starts from the first element in the image and it compares all the next element to calculate the similarity between them. This continues until all the elements along the row are clustered.

The column objects are then clustered using the coVAT algorithm using

$$[D_c]_{ij} = |d_i^* - d_j^*|, \text{ for } 1 \leq i \leq n, 1 \leq j \leq n.$$

The column object clustering also starts from the first element in the image and it compares all the next element to calculate the similarity between them. This continues until all the elements along the row are clustered.

The row and column objects are then combined. The iVAT algorithm is then applied to the combined matrix whish results in improved clustering image. The main advantage of using this technique is, it supports large datasets and provides more accuracy in clustering.

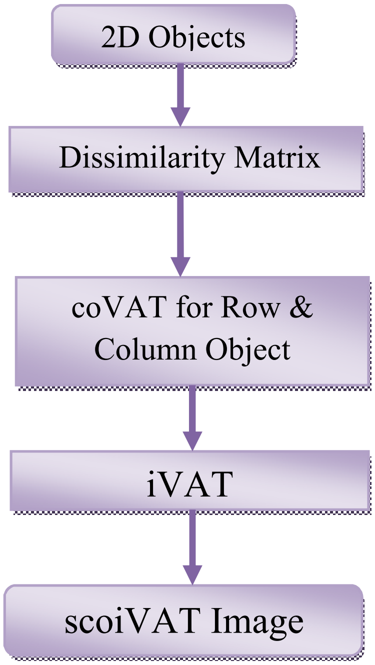

Fig.1.1.General architecture of clustering using scoiVAT

V. SYSTEM MODULES

A. 2D Object datasets

Getting 2D objects datasets as an input and these datasets are converted into dissimilarity matrix. Each data set is composed large number of 2D objects. 2D objects can be visualized using 2D data plots. The datasets contain x-coordinates and y-coordinates. These datasets are then converted into image that contains the plots of those coordinates.

B. Dissimilarity matrix

2D objects datasets image are converted into dissimilarity matrix using some distance transform algorithm. Mahalanobis distance calculation algorithm is used which correlates the x-coordinate and y-coordinate to produce the dissimilarity matrix. The dissimilarity matrix contains the values that represent each and every points specified in the plotted image.

C. *coVAT for row object*

Row object vector is taken as an input and clustering takes place along each row. This improves the clustering accuracy and build the cluster for row object using coVAT algorithm.

Row Matrix [Dr] is imputed using,
$[Dr]ij = |di,* - dj,*|$, for $1 \leq i \leq m$, $1 \leq j \leq m$.

D. coVAT for column object

Column object vector is taken as an input and clustering takes place along each column. This improves the clustering accuracy and build the cluster for column object using coVAT algorithm.

Column Matrix [Dc] is imputed using
$[Dc]ij = |di* - dj*|$, for $1 \leq i \leq n$, $1 \leq j \leq n$.

E. scoiVAT image

New matrix is obtained by combining both the row and column cluster elements Druc by extracting sampled rows and columns from coVAT algorithm. Then scoiVAT algorithm is applied on the new matrix. Final VAT cluster image is then displayed.

VI CONCLUSION

scoiVAT overcomes the problem of existing system and it provides efficient clustering method. scoiVAT is a scalable approach to four different cluster assessment problems associated with a very large M×N rectangular dissimilarity matrix DM×N. The problems are the assessment of cluster tendency: (P1) amongst the row objects; (P2) amongst the column objects; (P3) amongst the union of the row and column objects; and (P4) amongst the union of the row and column objects that contain at least one object of each type (coclusters). scoiVAT builds a sample Dn×n from DM×N , and then uses coVAT to find clustered images which overcomes the existing problems..

In the future, we will consider complexity and accuracy issues over clustering and also to check tendency and assessment of data objects.

REFERENCES

[1] J. Bezdek and R. Hathaway,(2002) “VAT: A Tool for Visual Assessment of (Cluster) Tendency,” Proc. Int’l Joint Conf. Neural Networks (IJCNN), pp. 2225-30.

[2] Enrique Pelayo and Carlos Orrite and David Buldain ,(2011)”SO-VAT: Self-Organizing Visual Assessment of cluster Tendency for large data sets” IEEE,vol 27-29.

[3] R. Hathaway, J. Bezdek, and J. Huband,(2006) “Scalable Visual Asseessment of Cluster Tendency for Large Data Sets,” Pattern Recognition, vol. 39, no. 7, pp. 1315-1324.

[4] T. Havens, J. Bezdek, J. Keller, and M. Popescu,(2009) “Clustering in Ordered Dissimilarity Data,” Int’l J. Intelligent Systems, vol. 24, no. 5, pp. 504-528.

[5] T. Havens, J. Bezdek, and J. Keller,(2010) “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.

[6] Isaac J. Sledge, Timothy C. Havens Jacalyn M. Huband James C. Bezdek James M. Keller (2009)” Finding the number of clusters in ordered dissimilarities”.

[7] Liang Wang, Senior Member, IEEE, Xin Geng, James Bezdek,(2010)“Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning” IEEE Trans. vol. 22, no. 10.

[8] Srinivasulu Asadi, Dr Ch D V Subba Rao, V Saikrishna,(2010)” Finding the Number of Clusters in Unlabeled Datasets using Extended Dark Block Extraction”.

[9] Timothy C. Havens, Senior Member,(2012)”An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm” IEEE, and James C. Bezdek, IEEE VOL. 24, NO. 5.

[10] Yingkang Hu,(2011)” VATdt: Visual Assessment of Cluster Tendency Using Diagonal Tracing” American Journal.