SURVEY ON VISUALIZED METHODS FOR DATA CLUSTERS ASSESSMENT

POLE ANJAIAH
Institute of Aeronautical Engineering
Email Id: anjaiah.pole@gmail.com

ABSTRACT
The problem of the access tendency plays a key role in cluster analysis. Numbers of clusters are need to be detected for effective cluster analysis; However, the existing visual access cluster tendency (VAT) method may not work effectively work on tight clustered data. Hence, the spectral procedure is used in VAT for achieving quality of cluster tendency results. In large data sets, it requires high computational time, since it depends on Eigen decomposition procedure. In this paper, an improved version of new clustering tendency method is proposed, which derives cluster tendency results through shortest paths. It is experimented on various synthetic clustered data sets for demonstrating the efficiency of proposed VAT method.

Keywords: *VAT, Cluster Analysis, Cluster Tendency Assessment, Spectral Approach.*

1. INTRODUCTION

The amount of data continues to grow at an enormous rate even though the data stores are already vast. The primary challenge is how to make the data base a competitive business advantage by converting seemingly meaningless data into useful information. How this challenge is met is critical because companies are increasingly relying on effective analysis of the information simply to remain competitive. A mixture of new techniques and technology is emerging to help sort through the data and find useful competitive data.

By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large databases thereby serve as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database systems and machine learning. Companies in many industries also take knowledge discovering as an important area with an opportunity of major revenue. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications.

Cluster analysis [9] has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city.

Groups of automobile insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery.

Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection, where outliers (values that are "far away" from any cluster) may be more interesting than common cases. Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce.

Effective and quality of clustering depends on cluster tendency assessment; there are many of cluster assessment methods are presented in [ ] for deriving of number of clusters. Visual access tendency is one of popular method for assessment of number of clusters. It works as efficient when the size of processed data is small and it is need to be enhanced for large tight clustered datasets. Thus, in proposed schema, spectral approach is used for handling of tight clustered datasets and shortest path mechanism is established for addressing the large data problems. Key contributions of the paper are summarized as follows:

1. Spectral approach is developed for addressing the tight clusters problems
2. Assessment of clusters are performed using primary idea of VAT method
3. Efficiency of proposed method is determined using some synthetic and bench-marked datasets.
4. Shortest path mechanism is established for finding similarities between data objects in a faster way in the proposed method
5. Proposed improved spectral based VAT (isVAT) is proposed for deriving of cluster tendency results of comprehensive datasets.

The proposed isVAT mechanism if effectively works on large and various bench-marked datasets for deriving of cluster tendency as well as cluster results. Remaining sections of the paper are described as follows: Section 2 overviews the related study, Section 3 describes the proposed work, Section 4 presents the experimental study, and Section 5 describes the conclusion and future scope.

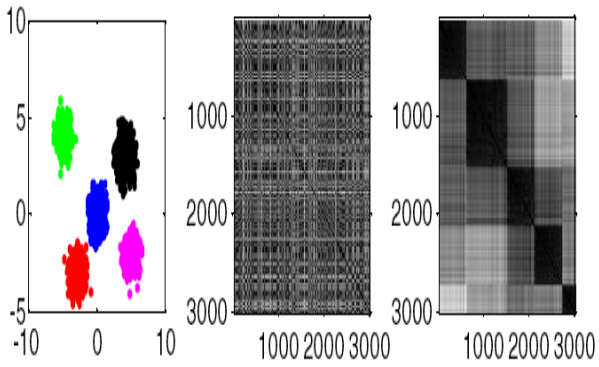
2. RELATED WORK

The problem of determining number of clusters prior to actual clustering is called the assessing ofclustering tendency. None of the existing approaches is completely satisfactory (nor will they ever be). The purpose of this note is to add a simple and intuitive visual approach to the existing repertoire of tendency assessment tools. The visual approach for assessing cluster tendency introduced here can be used in all cases involving numerical data. It is both convenient and expected that new methods in clustering have a catchy acronym. Consequently, we call this new tool VAT [ ]. The VAT approach presents pair wise dissimilarity information about the set of objects O = {o1, on} as a square digital image with n2 pixels, after the objects are suitably reordered so that the image is better able to highlight potential cluster structure.

To go further into the VAT approach requires some additional background on the types of data typically available to describe the set O = {o1… on}. The VAT is widely applicable because it displays a reordered form of dissimilarity data, which itself can always be obtained from the original data for 'O'. If the original data

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 'O'.

So, we can assume without loss that dissimilarity data of the type needed for a VAT display can be easily obtained, whether the original data description of 'O' is set of objects or rtelational and whether the data are complete or incomplete. Therefore, the VAT approach is applicable to virtually all numerical data sets.



Scatter Unord Reorder

Fig 1: Reordered VAT Image

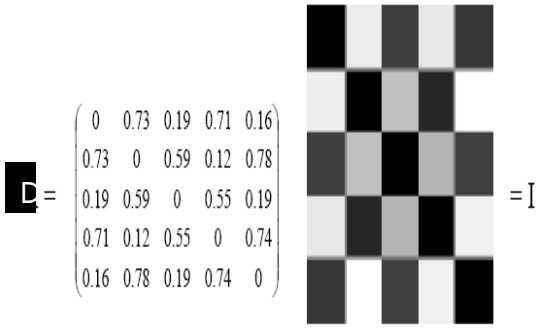


Fig 2: Dissimilarity matrix for Image

Algorithm VAT [ ]

Input: *Dissimilarity Matrix, D, NXN Matrix.*

1. Set I empty, j = {1,2,…n}, D = (0,0…0)

Select max value from D

P (1) = I;

I = { i }

J = J-{ i }

2. Repeat f or t = 2, 3… n

Select min value from $\{d_{pq}\}$*, where p is in I, q is in J*

Update I = I Ù {j}

J = J – {j}

3. Form the reordered matrix

D = d [p(i), p(j)]

In the above algorithm, step 1 shows the selection of objects with minimum similarity, step 2 updates the objects according to order of similarity features and step 3 shows the results of reordered dissimilarity matrix. It takes more computational time and other performance parameter values are degraded for large datasets. Thus, next section proposed isVAT method for determining of effective cluster tendency results of large datasets.

3. PROPOSED WORK

In proposed isVAT method, a weighted matrix is calculated the normalized Laplacian matrix is formed and finally RDI that portrays a potential cluster structure from the pair wise dissimilarity matrix of the data is created. For concreteness, we will generate RDIs using the Visual Assessment of Cluster Tendency (VAT) algorithm. Then, sequential image processing operations(region segmentation, directional morphological filtering and distance transformation) are used to segment the regions of interest in the RDI and to convert the filtered image into a distance-transformed image. Finally, we project the transformed image onto the diagonal axis of the RDI, which yields a one-dimensional signal, from which we can extract the (potential) number of

clusters in the data set using sequential signal processing operations like average smoothing and peak detection.

The proposed method is easy to understand and implement, and encouraging results are achieved on a variety of artificially generated and real-world data sets,we review related work and we describe the DBE approach that contains a description about the Spectral analysis. Here, we compare Spectral VAT to a predecessor algorithm called Dark Block extraction algorithm pointing out similarities and differences between the two approaches

Algorithm: isVAT

1. Compute the local scaling parameter oi for object Oi

2. Construct the weighting matrix W

3. Construct the normalized Laplacian matrix L

4. Choose the k largest eigenvectors of L' to form the matrix V

5. Normalize the rows of V with unit Euclidean norm to generate V'

6. Construct a new pair wise dissimilarity matrix D'

7. Apply the VAT algorithm to D'

The spectral decomposition of the Laplacian matrix provides useful information about the properties of the graph. It has been shown experimentally that natural groups in the original data space may not correspond to convex regions, but once they are mapped to a spectral space spanned by the eigenvectors of the Laplacian matrix, they are more likely to be transformed into tight clusters. Based on this observation, we wish to embed D in a k-dimensional spectral space, where k is the number of eigenvectors used, such that each original data point is implicitly replaced with a new vector instance in this new space. After a comprehensive study of recent spectral methods, we adopt a combination of adjacency graph, weighting function, and graph Laplacian for obtaining a better graph embedding.

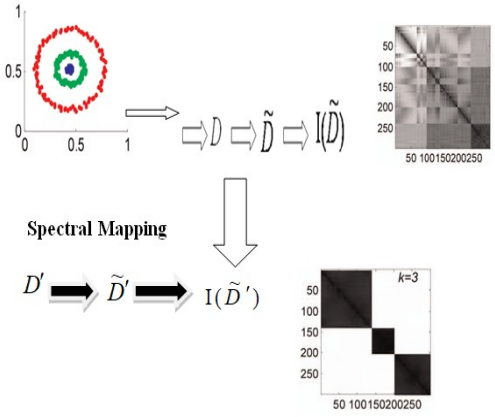


Fig 3: Spectral Mapping

4. EXPERIMENTAL STUDY

To enable automatic determination of the number of clusters, we need to find a best Spectral VAT image in terms of clarity and block structure. Each of the block regions in the image corresponds to either inter cluster or intra cluster dissimilarity values, while the clarity is relevant to the degree of the brightness difference between such blocks. The corresponding gray scale histograms suggest that a good Spectral VAT image should include two explicit modalities in the gray scale histogram, with a narrow distribution of each modality and a large distance between the two modalities. It is easily to understand that the toe modalities in the histogram implicitly correspond to with-in cluster distances and between cluster distances. A narrow distribution for any one modality means that values in either with-in cluster distances or between-cluster distances are close, where as a big distance between two modalities means that these two modalities are easily distinguished.

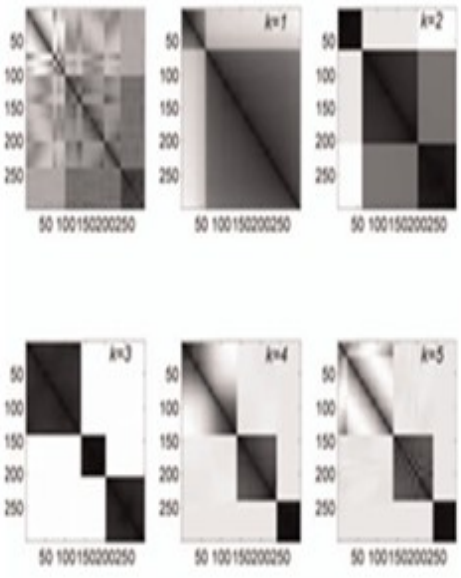


Fig 4: Cluster Tendency Assessment

Two Clustered data and VAT image for two clustered data

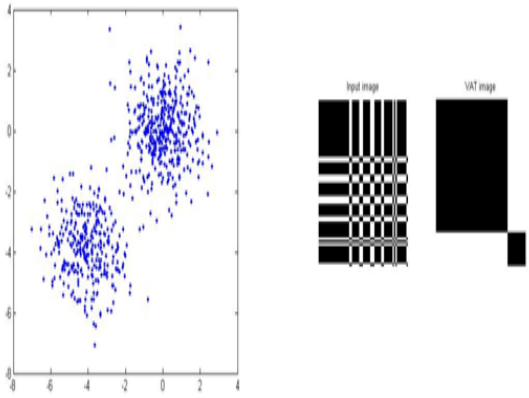


Fig 5: Results of Two Cluster Data Assessment

We are interested in further exploring the use of image based approaches to assess cluster tendency and extract information about the geometric structure of possible clusters. Questions of interest include finding alternative, superior ordering methods. The method here is closely related to clustering. Cutting the longest connecting edges in produces exactly the single linkage clusters for c = 2 in this data.

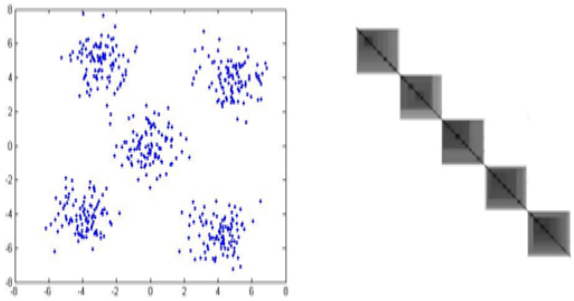


Fig 6: Results of Five Cluster Data Assessment

Five clusters in the data set are visually apparent but there is a high level of mixing between outliers from components in the mixture. Computing squared Euclidean distance between the pairs of vectors yields a matrix D with dissimilarities for the data set. Accessing only the vectors needed to make a particular distance computation and releasing the memory used by the vectors, to avoid the exhausting memory the processing was broken up calling the extension routine multiple times. We set the number of clusters C=5 for the data set, since running this data set took quite long time the appearance of error rates are small, it is not an easy clustering problem in terms of how well separated the clusters actually are.

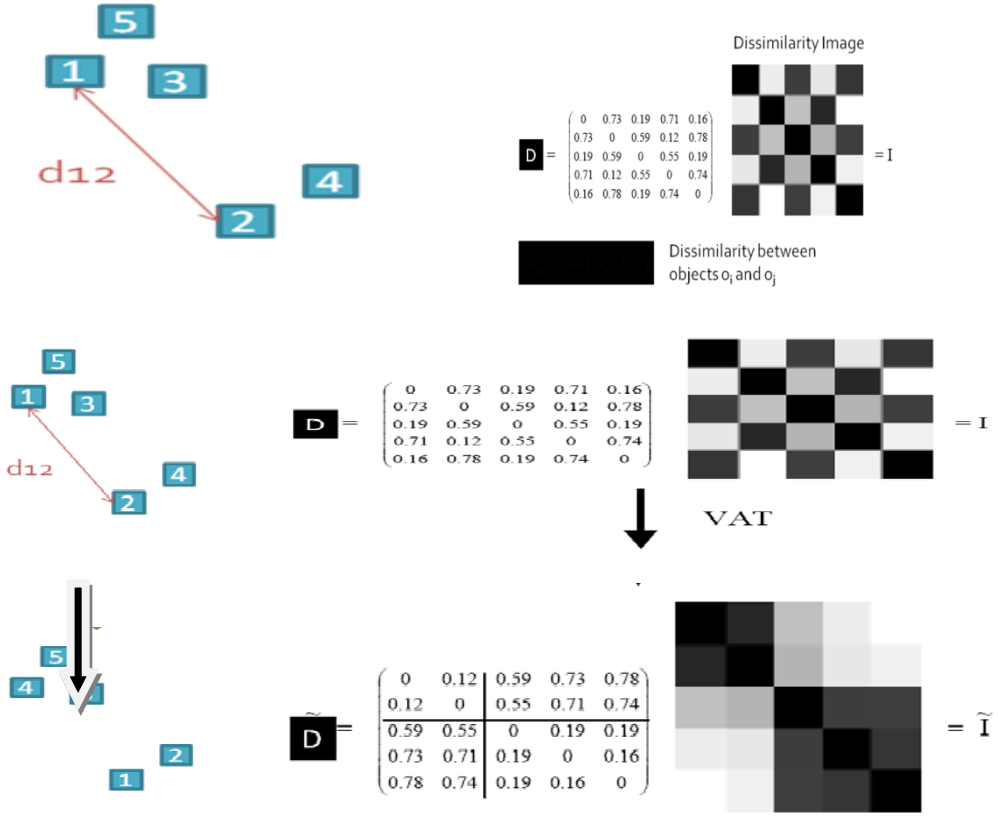


Fig 7: Results of isVAT algorithm for Sample Example.

5. CONCLUSION AND FUTURE SCOPE

Visual access tendency method is generally used for assessing the tendency or clusters of comprehensive datasets. For large datasets, it is enhanced using spectral concept for improving the cluster assessment results. This paper is presented an enhanced visual approach towards automatically determining the number of clusters and partitioning data in either object or pair wise relational form, to better reveal the hidden cluster structure, especially for complex-shaped data sets, the VAT algorithm has been improved by using spectral analysis of the proximity matrix of the data. Based on Spectral VAT, a goodness measure of Spectral VAT images has been proposed for automatically determining the number of clusters, derived a visual clustering algorithm based on Spectral VAT images and its unique blocked structured property, and also proposed an extended strategy to scale the Spectral VAT algorithm to larger data sets. Future scope of the proposed method is to be extended for assessment of video cluster tendency results which useful for effective video surveillance results.

REFERENCES

[1] *Data Mining: Concepts and Techniques (2nd Edition) by " Han and Kamber"*

[2] *L.Waing, C.Leckie and J.C.Bezdek, "Automatically Determining the Number of Clusters in Unlabeled Datasets", IEEE Transaction on Knowledge and Data engineering, vol. 21, no. 3, 2009.*

[3] *T.Havens, J.C.Bezdek, J.Keller and M.Popesu, Dunn.s "Cluster Valid index as a Contrast Measure of VAT Images", IEEE, 2008.*

[4] *L.Waing ,Geng , J.Bezdek and C.Leckie , .Enhanced Visual Analysis for Cluster tendency assessment and Data Partitioning. , IEEE Transaction on Knowledge and Data engineering , vol.22, no.10, pp.1401-1414, 2010.*

[5] *L.Waing, C.Leckie and J.C.Bezdek, .Automatically Determining the Number of Clusters in Unlabeled Datasets., IEEE Transaction on Knowledge and Data engineering, vol. 21, no. 3, 2009.*

[6] *T.Havens, J.C.Bezdek, J.Keller and M.Popesu, .Dunn.s Cluster Valid index as a Contrast Measure of VAT Images., IEEE, 2008.*

[7] *SanghamitraBandyopadhyay and Sripamasaha, .A Point Symmetry based Clustering Technique for Automatic Evolution of clusters., IEEE Transaction on Knowledge and Data engineering, vol. 20, no. 11, 2008.*

[8] *I.Sledge, J.Huband and J.C.Bezdek, .(Automatic) Cluster Count Extraction from Unlabeled Datasets., Joint Proceedings Fourth Int.l Conference Natural Computation (ICNC) and Fifth Int.l Conference on Fuzzy Systems and Knowledge discovery (FSKD), 2008.*

[9] *J.C.Bezdek, R.J.Hathway and J.Huband, .Visual Assessment of Fuzzy Clustering Tendency for Rectangular Dissimilarity Matrices., IEEE Transactions on Systems, vol. 15, no. 5, pp. 890-903, 2007.*

[10] *L.Waing and Y. Zhang , .On fuzzy cluster validity indices., Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2095-2117, 2007.*

[11] *R.Hathway, J.C.Bezdek and J.Huband, .Scalable Visual Assessment of Cluster Tendency., Pattern Recognition, vol.39, no. 6, pp. 1315-1324, 2006.*

[12] *J.Huband, J.C.Bezdek and R.Hathway, .BigVAT: Visual Assessment of Cluster Tendency ., Pattern Recognition, pp. 1875-1886, 2005.*

[13] *GautamGarai, B.B.Chaudhuri, .A Novel Genetic Algorithm for Automatic Clustering., Pattern Recognition, Science Direct Letters 25, pp. 173.187, 2004.*

[14] *U.Maulik and S. Bandyopadhyay, .Performance Evaluation of Some Clustering Algorithms and Validity Indices., IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, 2002.*

[15] *J.C.Bezdek and N.R.Pal, .Some new indexes of cluster validity., IEEE Transactions on Systems, Man, And Cybernetics, vol. 28, pp. 301.315, 1998*

[16] *J.G.Milligan and M.Cooper, .An Examination of Procedures for Determining the Number of Clusters in a Data Set., Psychometrika, vol. 50, pp. 159-179, 1985.*

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

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

[19] *P.Sneath, .A Computer Approach to Numerical Taxonomy., J. General Microbiology, vol. 17, pp. 201-226, 1957.*