VAT: A tool for visual assessment of (cluster) tendency

Conference Paper · February 2002

DOI: 10.1109/IJCNN.2002.1007487 · Source: IEEE Xplore

2 authors, including:

James C. Bezdek
University of Missouri
408 PUBLICATIONS **58,132** CITATIONS

SEE PROFILE

VAT: A Tool for Visual Assessment of (Cluster) Tendency

J.C. Bezdek*
Department of Computer Science
University of West Florida
Pensacola, FL 32514

R.J. Hathaway*
Mathematics and Computer Science Department
Georgia Southern University
Statesboro, GA 30460

***Abstract*-A method is given for visually assessing the cluster tendency of a set of Objects O = {o$_1$,…,o$_n$} when they are represented either as object vectors or by numerical pairwise 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.**

I. Introduction

We consider a type of preliminary data analysis related to the pattern recognition problem of clustering. Clustering or cluster analysis is the problem of partitioning a set of objects O = {o_1,…, o_n} into c self-similar subsets based on available data and some well-defined measure of (cluster) similarity. In some cases, a geometric description of the clusters (e.g. by "cluster centers" in data space) is also desired and some clustering methods are able to produce such geometric descriptors. The type of clusters found is strongly related to the properties of the mathematical model that underlies the clustering method. All clustering algorithms will find an arbitrary (up to $1 \le c \le n$) number of clusters, even if no "actual" clusters exist. Therefore, a fundamentally important question to ask before applying any particular (and potentially biasing) clustering algorithm is: Are clusters present at all?

The problem of determining whether clusters are present as a step prior to actual clustering is called the *assessing of clustering tendency*. Various formal (statistically based) and informal techniques for tendency assessment are discussed in Jain and Dubes [1] and Everitt [2]. 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.

Visual approaches for various data analysis problems have been widely studied in the last 25 years; Tukey [3] and Cleveland [4] are standard sources for many visual techniques. 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 (*visual assessment of tendency*). The VAT approach presents pair wise dissimilarity information about the set of objects O = {o_1,…, o_n} as a square digital image with n^2 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 = {o_1,…, o_n}.

There are two common data representations of O upon which clustering can be based. When each object in O is represented by a (column) vector x in $\Re^s$, the set X = {x_1,…,x_n} $\subset \Re^s$ is called an *object data* representation of O. The k^{th} component of the i^{th} feature vector (x_{ki}) is the value of the k^{th} feature (e.g., height, weight, length, etc.) of the i^{th} object. It is in this data space that practitioners sometimes seek geometrical descriptors of the clusters. Alternatively, when each *pair* of objects in O is represented by a relationship, then we have *relational data*. The most common case of relational data is when we have (a matrix of) dissimilarity data, say R = [R_{ij}] , where R_{ij} is the pair wise dissimilarity (usually a distance) between objects o_i and o_j, for $1 \le i, j \le n$. More generally, R can be a matrix of similarities based on a variety of measures [5,6].

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 O. If the original data consists of a matrix of pair wise (symmetric) similarities S = [S_{ij}], then dissimilarities can be obtained through several simple transformations. For example, we can take

$$R_{ij} = S_{max} - S_{ij}, \qquad (1)$$

where S_{max} denotes the largest similarity value. If the original data set consists of object data X = {x_1,…,x_n} $\subset \Re^s$, then R_{ij} can be computed as $R_{ij} = \left\| x_i - x_j \right\|$, using any convenient norm on $\Re^s$.

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. Consequently, sophisticated imputation schemes, such as those based on the expectation-maximization (EM) algorithm in Dempster, Laird and Rubin [7], are unnecessarily expensive in both complexity and computation time. For incomplete object data, we would suggest the Dixon [8] scheme, which generates a pair wise Euclidean (or other norm) dissimilarity R_{ij} from incomplete x_i and x_j 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. For missing dissimilarity values (R_{ij}), one of the triangle inequality schemes in Hathaway and Bezdek [9] should be sufficiently accurate. We refer the reader interested in learning more about missing data and imputation to Little and Rubin [10] and Schafer [11].

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 object or relational, and whether the data are complete or incomplete.

* Research supported by ONR Grant 00014-96-1-0642.

Therefore, the VAT approach is applicable to virtually *all* numerical data sets. In the next section we define the main idea of the VAT approach and then give a description of how it can be implemented. Section 3 discusses relatives of VAT, while Section 4 gives a series of examples using various real and artificial data sets that illustrate various facets of the VAT tool. We attempt to give the reader a feel for what various types of cluster structure (including no cluster structure) may look like in VAT displays. The final section contains some concluding remarks and topics for further research.

II. Ordered Dissimilarity Images

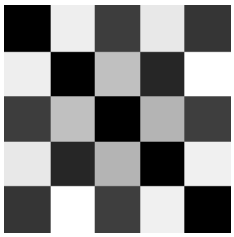
Let R be an $n \times n$ dissimilarity matrix corresponding to the set $O = \{o_1,\ldots, o_n\}$. We assume that R satisfies the following (metric) conditions for all $1 \le i, j \le n$:

$$R_{ij} \ge 0 \quad (2a)$$
$$R_{ij} = R_{ji} \quad (2b)$$
$$R_{ii} = 0 \quad (2c)$$

We display R as an intensity image I, which we call a *dissimilarity image*. The intensity or gray level g_{ij} of pixel (i,j) depends on the value of R_{ij}. The value $R_{ij} = 0$ corresponds to $g_{ij} = 0$ (pure black); the value $R_{ij} = R_{max}$, where R_{max} denotes the largest dissimilarity value in R, gives $g_{ij} = R_{max}$ (pure white). Intermediate values of R_{ij} produce pixels with intermediate levels of gray in a set of gray levels $G = \{G_1,\ldots,G_m\}$. The images shown below use 256 equally spaced gray levels, with $G_1 = 0$ (black) and $G_m = R_{max}$ (white). The displayed gray level of pixel (i,j) is the level $g_{ij} \in G$ that is closest to R_{ij}.

As an example, Fig. 1 lists a small dissimilarity matrix and its corresponding image. The 0 values on the main diagonal of R generate main diagonal pixels that are black. Notice also that the largest dissimilarity value (0.78) gives two white pixels in the dissimilarity image.

$$R = \begin{pmatrix} 0 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0 \end{pmatrix}$$



$= I$

Fig. 1. A dissimilarity matrix and its corresponding dissimilarity image.

Does the image in Fig. 1 indicate that clusters are likely for the five objects underlying the relational data shown there? More generally, can a dissimilarity image indicate the presence of clusters? We surmise that the usefulness of a dissimilarity image for visually assessing cluster tendency depends crucially on the *ordering* of the rows and columns of R. We will attempt to reorder the objects $\{o_1,o_2,\ldots,o_n\}$ as $\{o_{k_1}, o_{k_2}, \ldots, o_{k_n}\}$ so that, to whatever degree possible, if k_i is near k_j, then o_{k_i} is similar to o_{k_j}. In this case, the corresponding *ordered dissimilarity image* (ODI) $\tilde{I}$ will often indicate cluster tendency in the data by dark *blocks* of pixels along the main diagonal. The ordering is accomplished by processing elements in the dissimilarity matrix R (rather than using the objects or object data directly). A procedure for ordering is given immediately after our second example.

Fig. 2 is a scatterplot of n = 20 points in a 2-dimensional data set, called Data Set A, that we use to illustrate the importance of properly ordering the rows and columns of the dissimilarity matrix for visual assessment of tendency. This data set has (either) three visually apparent clusters and one outlying point, or 4 clusters, if singleton clusters are allowed. Fig. 3 shows the 19 sequential distances $\{d_{12}, d_{23}, \ldots, d_{19,20}\}$ between points $\{x_1, x_2, \ldots, x_{20}\}$, where indices $\{1,\ldots,20\}$ correspond to a random initial ordering of the points. The ordering of the data in the scatterplot is indicated by the path of line segments, with the firstly ordered point represented by the heavy square. The corresponding dissimilarity image in the right view of Fig. 3 contains no useful (visual) information about (apparent) structure in Data Set A. Now, we reorder the points so that nearby points are (generally) indexed similarly. Fig. 4 gives the scatterplot for the reordered points along with the corresponding ordered dissimilarity image. The ODI indicates the likelihood of clusters, as seen by the one large and 2 smaller blocks of dark pixels along the main diagonal of the ODI. The isolated outlier is seen as the single black diagonal pixel in the last row and column. The line segments shown in the left view of Fig. 4 indicate the sequence of (reordered) indices $\{k_1, \ldots, k_{20}\}$, with x_{k_1} again indicated by the heavy dark square.

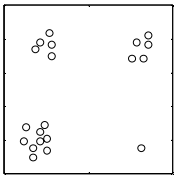


Fig. 2. Scatterplot of Data Set A.

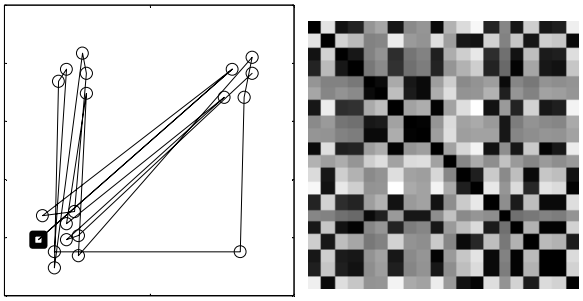


Fig. 3. Scatterplot and dissimilarity image for Data Set A (original random ordering).

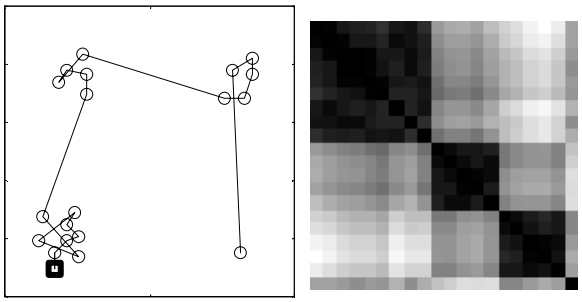


Fig. 4. Scatterplot and ODI for Data Set A (reordered)

The dark diagonal blocks in the ODI of Figure 4 clearly indicate the presence of 1 large and 2 smaller clusters, as well as the isolated singleton in data set A. The mechanism underlying the emergence of visually clear *blocks* on the diagonal of the display is simple. If we order the points so that nearby points (generally) have similar index values, then rows of R with similar index values will be similar. This will give repetition in the pixel patterns of nearby groups of rows of R, which will in turn give rise to a visible block structure in the ODI. A black block corresponds to a set of nearby points, consecutively ordered. Without the proper ordering, it is essentially impossible to visually assess clustering tendency using a dissimilarity image. We will now state one possible algorithm for suitably ordering the rows and columns of the dissimilarity matrix.

Our ordering algorithm is similar to Prim's algorithm for finding a *minimal spanning tree* (MST) of a weighted graph. (See, for example, [12], for a description of Prim's algorithm.) The main differences between our algorithm and Prim's algorithm are that: (i) we are not interested in representing the MST, but only in finding the order in which the vertices are added as it is grown; and (ii), we specify a method for choosing the initial vertex that depends on the maximum edge weight in the underlying complete graph.

It is convenient to introduce some helpful notation. Let I and J be subsets of K = {1, …, n}. We let $\arg\min_{p\in I, q\in J} \{R_{pq}\}$ denote the set of *all* ordered index pairs (i,j) in $I \times J$ such that $R_{ij} = \arg\min_{p\in I, q\in J} \{R_{pq}\}$. This differs from the usual meaning of "arg min" only in that a call to arg min (f(*)) ordinarily returns only one value of (*) that minimizes f, whereas here we collect *all* values of (*) that yield the (same) minimizing value. The "arg max" notation is defined similarly. The algorithm for producing an ordered dissimilarity matrix $\tilde{R} = [R_{k_i k_j}]$ from the original dissimilarity matrix R is now given. The permuted indices of the n objects are stored in an array P[ ], with $P(i) = k_i$, i = 1, …, n.

VAT Ordering and Display Algorithm

Step 1 Set K = {1,2,…,n}; I = J = ∅; P[0] = (0,…,0).

Step 2 Select $(i,j) \in \arg\max_{p\in K, q\in K} \{R_{pq}\}$.
Set P(1) = i; I = {i}; and J = K – {i}.

Step 3 For r = 2,…, n:
Select $(i,j) \in \arg\min_{p\in I, q\in J} \{R_{pq}\}$.
Set P(r) = j; Replace I ← I∪{j} and J ← J–{j}.
Next r.

Step 4 Obtain the ordered dissimilarity matrix $\tilde{R}$ using the ordering array P as: $\tilde{R}_{ij} = R_{P(i)P(j)}$, for $1 \le i,j \le n$.

Step 5 Display the reordered matrix $\tilde{R}$ as the ODI $\tilde{I}$ using the conventions given above.

We make a few comments. For convenience, think of the objects as ordered points in space (e.g., an object data representation of the objects.) First, note that distances in $\tilde{R}$ are *not* recomputed; instead, we are simply rearranging the rows (and columns) of R. Step 2 of the VAT algorithm finds the set of longest edges in R. This choice initiates the ordering at one of the "outermost" points of the data. The intended purpose of this initialization is to avoid beginning at an "interior" point, where the potential for "zigzagging" between (possible) clusters while ordering is greatest. Cluster zigzagging, illustrated in Figure 13, can detrimentally affect the display of diagonal blocks. Tie breaking strategies must be imposed as there will always be at least two choices (viz., either vertex at the ends of a longest edge) for $P(1) = k_1$ in Step 2, and possibly multiple choices for $P(r) = k_r$ in Step 3. Finally, we mention that another way to handle missing components of R is to simply constrain the index searches in Steps 2 and 3 over only non-missing elements. Step 5 simply displays $\tilde{R}$ as the ODI $\tilde{I}$.

The VAT algorithm is applied to the example of Fig. 1 and the results are displayed in Fig. 5.

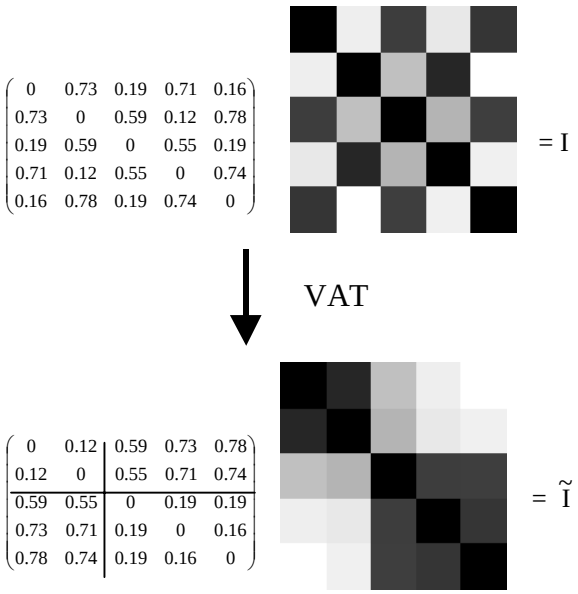


Fig. 5. Results of applying the VAT algorithm to Data Set A.

The lower view in Fig. 5 contains the pair $(\tilde{R}, \tilde{I})$. Our inference from $\tilde{I}$ is that the objects underlying R lie in two small clusters, one having two objects, the other with three. Examination of $\tilde{R}$ confirms this, as it possesses two diagonal blocks of sizes 2×2 and 3×3 which evidently correspond to the substructure visually suggested by $\tilde{I}$.

III. RELATIVES OF VAT

We can roughly group visual display methods into three categories: visual displays *of* clusters, visual displays *to find clusters*, and visual displays *to assess tendency*. (Admittedly, there is a very fine line distinguishing methods in the second and third groups.) The earliest published reference we can find that discusses visual displays (as images) of clusters is the 1973 SHADE approach of Ling [13]. SHADE approximates what is now regarded as a nice digital image

representation of clusters using a crude 15 level halftone scheme created by overstriking standard printed characters. SHADE displays the lower triangular part of the complete square display. Visual identification of (triangular) patterns is considerably more difficult than when a full, square display is used. SHADE was used *after* application of a hierarchical clustering scheme, as an alternative to visual displays of hierarchically nested clusters via the standard dendrogram.

Closely related to SHADE, but presented more in the spirit of finding clusters (i.e., as a visual clustering algorithm) rather than displaying clusters found with an outsourced algorithm is the "graphical method of *shading*" described on p. 577 of [18]. Johnson and Wichern give this informal description: (i) arrange the distances into several classes of 15 or fewer, based on their magnitudes; (ii) replace all distances in each class by a common symbol with a certain shade of gray; (iii) reorganize the distance matrix so that items with common symbols appear in contiguous locations along the main diagonal (darker symbols correspond to smaller distances); (iv) groups of similar items correspond to patches of dark shadings. The example presented in this text is remarkably similar to the 1973 SHADE image: triangular, and made with groups of symbols. Step (i) in this method amounts to finding clusters with an ad hoc criterion, and so, differs from SHADE in the way that clusters are produced. The method of reordering in step (iii) is based on the clusters found, and so, is somewhat different from VAT.

Various writers have made image displays of distance data. See, for example, the demo found under GENLAB at the website [17], which displays (unordered) distance matrices as color images. We could not make the GENLAB software reorder the distances to create a display like the VAT tool. However, images of this type might be used to make guesses about cluster tendency. We feel that displays such as these of the unordered distance matrix will not be particularly useful.

IV. Numerical Examples

We expect dark, block structure along the main diagonal of the ODI for data sets containing well-separated clusters (such as those in Fig. 3). As the degree of separation between clusters decreases, the clarity of dark diagonal blocks diminishes. To illustrate this, we will use normal mixture data sets similar, to those used in Bezdek and Pal [14] to study various cluster validity schemes. (Cluster validity is the post-clustering problem of determining if a particular computed clustering is consistent with the data.) We describe how the normal mixtures are generated.

Let e_i denote the i^{th} unit vector in $\Re^4$ and I_4 denote the 4 x 4 identity matrix. First, we generate a set of 128 4-dimensional observations from a multivariate normal distribution having mean vector $(0,0,0,0)^T$ and covariance matrix I_4. The 128 observations are then divided into 4 groups of 32 each (so c = 4). For $1 \le i \le 4$, the i^{th} group is altered by adding α to the i^{th} component of each of the 32 points. (This is statistically equivalent to generating the i^{th} group of observations from a multivariate normal distribution with mean $\mu_i = \alpha e_i$ and covariance matrix I.) In all, we use the 5 values α = 4, 3, 2, 1, and 0 to get 5 data sets, each derived from the original 128 points. We denote the normal data set corresponding to a given value of α by Normal(α).

For example, the distributional means of the sample Nomal(4) are: $\mu_1 = (4,0,0,0)^T$; $\mu_2 = (0,4,0,0)^T$; $\mu_3 = (0,0,4,0)^T$; and $\mu_4 = (0,0,0,4)^T$. The means for Normal(1) are just e_1, e_2, e_3 and e_4. And when α = 0, Normal(0) is a single Gaussian cluster centered at the origin of $\Re^4$. For each data set Normal(α), we use the same random reordering, and then display the random and ordered dissimilarity image in Figs. 6-10 below. We expect Normal(4) to show some clear signs of cluster tendency; and we expect the visual evidence for c = 4 clusters to deteriorate as α decreases. Relational data is derived as Euclidean distances from the object data.

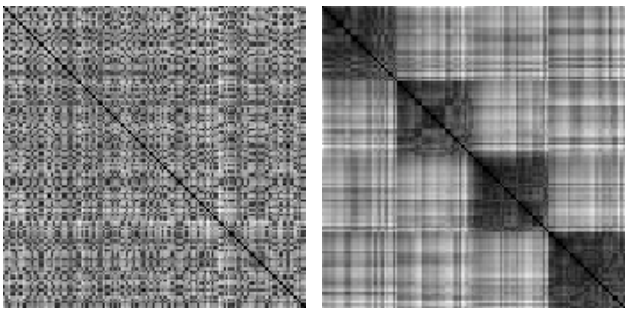


Fig. 6. Random and ordered dissimilarity images for Normal(4) data.

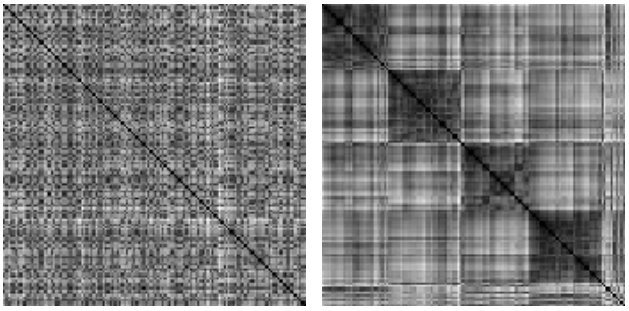


Fig. 7. Random and ordered dissimilarity images for Normal(3) data.

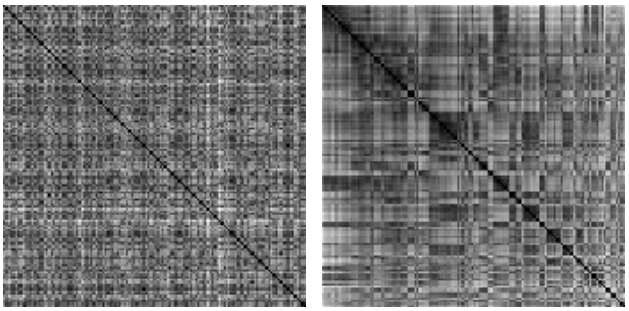


Fig. 8. Random and ordered dissimilarity images for Normal(2) data.

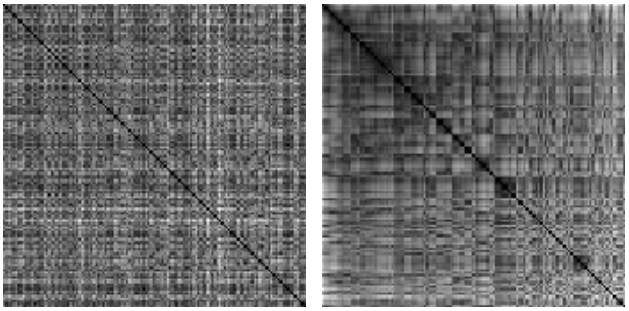


Fig. 9. Random and ordered dissimilarity images for Normal(1) data.

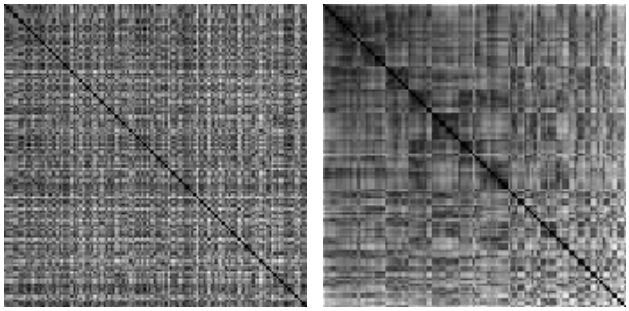


Fig. 10. Random and ordered dissimilarity images for Normal(0) data.

Figs. 6-10 give some information about the degree of separation needed for cluster-indicating blocks to appear on the diagonal of the ODI. The blocks are clear for $\alpha = 4$ and 3, and deteriorate dramatically for $\alpha = 2$. ODI's of the type in Figs. 8-10 are typical of data sets that do not have well separated clusters.

We next give an ordered dissimilarity image for the ubiquitous IRIS data. We use the "real" version of the data as discussed in [15]. The original object data consists of measurements of 4 features of each of 150 irises. The 150 irises are of three different physical types, and 50 of each type were used to generate the original object data. Thus, the IRIS data has three physically labeled classes. However, it is well known that two of the three flower types yield data that greatly overlap in $\Re^4$, so it is often argued that the unlabeled data are naturally clustered into 2 (geometrically well-defined) clusters. Relational data, randomly ordered, were generated as pair wise Euclidean distances between each pair of object data. Fig. 11 gives the random and ordered dissimilarity images for the IRIS data. The ODI indicates that there are 2 well-separated clusters. This assessment agrees with our intuition, and with results obtained in other ways in earlier published studies [19].

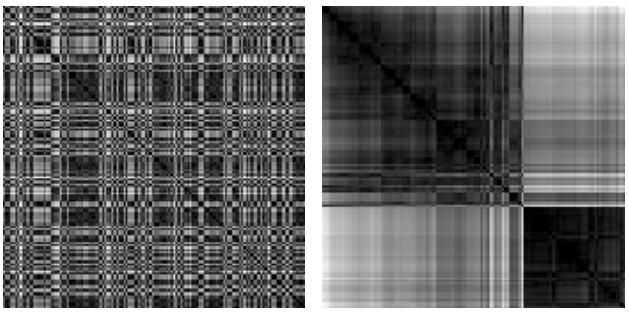


Fig. 11. Random and ordered dissimilarity images for the IRIS data.

Our next example uses a small set of real data from Gowda and Diday[16]. These data are obtained by applying a *similarity* measure to 5-dimensional object data, which has four quantitative, and one nominal qualitative feature values for each of eight different types of oil. The dissimilarity data, obtained by converting the similarity data from [16] to dissimilarity data using (1) and then randomly reordering, is given by

$$R_{FO} = \begin{pmatrix} 0 & 1.555 & 1.760 & 0.375 & 1.580 & 0.550 & 1.150 & 0.380 \\ 1.555 & 0 & 0.130 & 1.280 & 2.890 & 1.515 & 1.960 & 1.160 \\ 1.760 & 0.130 & 0 & 1.445 & 3.070 & 1.695 & 2.110 & 1.360 \\ 0.375 & 1.280 & 1.445 & 0 & 1.670 & 0 & 0.650 & 0.160 \\ 1.580 & 2.890 & 3.070 & 1.670 & 0 & 1.615 & 1.010 & 1.880 \\ 0.550 & 1.515 & 1.695 & 0 & 1.615 & 0 & 0.560 & 0.315 \\ 1.150 & 1.960 & 2.110 & 0.650 & 1.010 & 0.560 & 0 & 0.850 \\ 0.380 & 1.160 & 1.360 & 0.160 & 1.880 & 0.315 & 0.850 & 0 \end{pmatrix}$$

Fig. 12 gives the random and ordered images for R_{FO}. The ODI indicates clusters. The clustering approach in [16] divides the 8 types of oils into c = 3 clusters. Specifically, the first 2 rows (or columns) correspond to C_1 = {Beef-tallow, Lard}; rows 3-6 correspond to C_2 = {Olive oil, Cotton-seed, Sesame oil, Camellia}; and the last 2 rows give C_3 = {Perilla oil, Linseed oil}. A single-linkage dendrogram, cut at the appropriate level, gives the two clusters corresponding, respectively, to the first two rows and the last 6 rows of the ODI in Fig. 12.

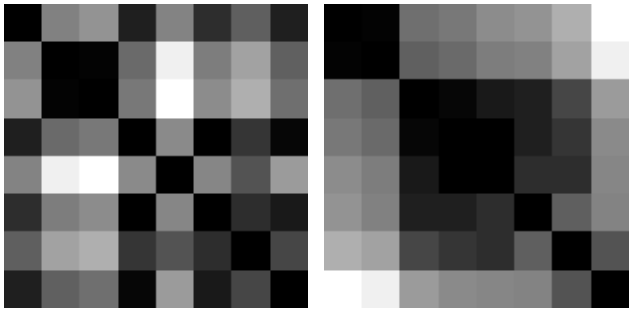


Fig. 12. Random and ordered dissimilarity images for the Fat-Oil data [16].

The next examples hint at the potential of VAT to extract some useful information about the geometry of object data clusters underlying the corresponding relational data. In Figs. 13-14, we show the results of applying VAT to linear and concentric circles grids of 128 data points, respectively. The data set underlying the three views in Fig. 13 consists of 8 rows of closely spaced points, 16 points per row. The upper view in Fig. 13 shows the image corresponding to a display of the unordered distances, with the initial point again being indicated by the dark circle (point number 3, left to right, uppermost row). The middle view in Fig. 13 is the ODI made by VAT, with the initial point being the first point (uppermost left) in the data. Here you can see that the data are traversed in a fully ordered walk of minimum distance. The bottom view in Fig. 13 illustrates the "zigzagging" effect that can result if the initialization for reordering made at Step 2 of our VAT algorithm is ignored. In this view, we chose a random initial vertex (the fifth point from the right in row 5 of the data). The resultant ODI still has the cyclic structure evident in the middle view, but there is also a c = 2 block structure superposed on it which confuses the viewer, and obscures possible interpretations of tendencies in the data.

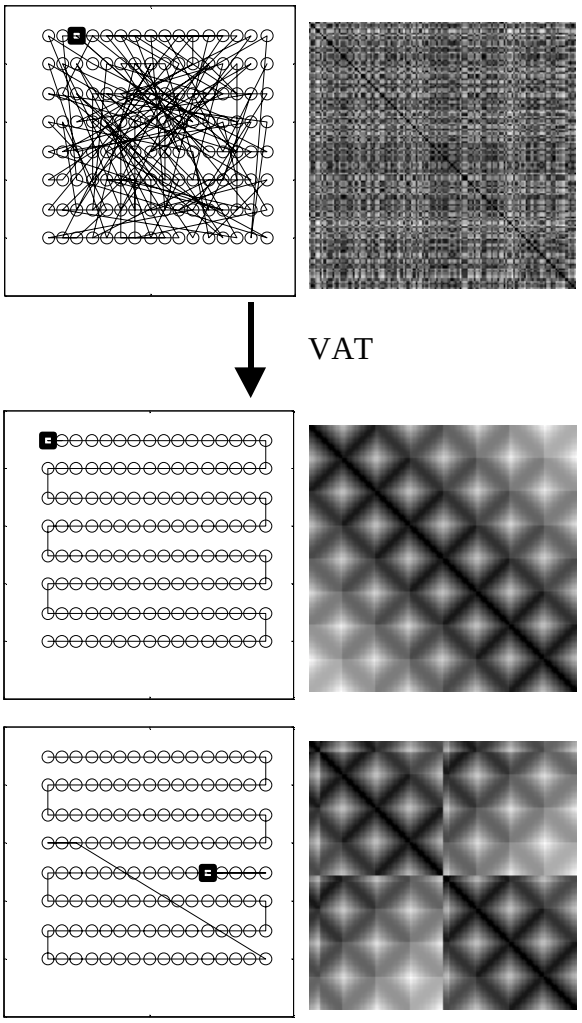


Fig. 13. Random and ordered dissimilarity images for linear data.

The data for Fig. 14 are drawn uniformly from a pair of concentric circles. This type of data set occurs often in real images, and has long posed a challenge for researchers in clustering. An interesting question that naturally arises after examining Figs. 13-14 is whether or not it is possible to correctly interpret the geometric alignment of object data based on ordered dissimilarity images. We believe that this can be done somewhat. For example, the ODI's in these figures indicate that both object data sets are arranged in a very regular fashion. In the ODI of Fig. 13, we can conclude the periodic nature of the object data arrangement because of the cycling from dark to light to dark, etc., as we go from left to right across a particular line of the image. Moreover, you can count 8 diagonal blocks, so it is reasonable to conjecture that the underlying data fall into 8 very regular clusters (which they do).

We can also infer that the data set corresponding to the ODI in Fig. 14 consists of 2 similar, regular structures, based on the 2×2 block form of the ODI image. Further examination of each diagonal block shows (by scanning a line in the block from left to right) that the objects are arranged in some type of loop, as the distances go through one cycle of varying levels of nearness. Of course, real data sets will not be this regular, but we still have some hope that certain geometric properties may manifest themselves in an ODI in a form that is recognizable.

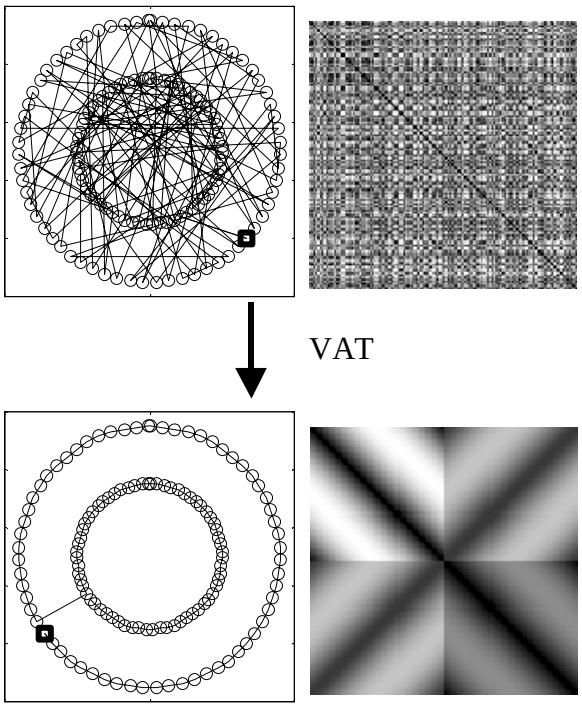


Fig. 14. Random and ordered dissimilarity images for circular data.

V. CONCLUSIONS

We gave a new approach for visually assessing cluster tendency using ordered dissimilarity images. The proposed ordering algorithm is related to Prim's algorithm for finding the minimal spanning tree of a weighted graph. The approach is able to signal the presence of well separated clusters via the manifestation of dark blocks of pixels on the main diagonal of the ODI. This technique is applicable to all dimensions and all numerical data types, complete or incomplete. Several 2 dimensional examples suggest that ODI's may help us "see" geometric properties of underlying object data sets.

An implementational issue is worth mentioning here. In order to get results like ours in Figs. 12 and 13, the tie breaking strategies implicit in VAT should favor additions of new points that are of minimal distance to the *most recently added* points. We illustrate this with an example. Suppose we have determined x_{k_1}, x_{k_2}, and x_{k_3} and that the minimum distance between any of our determined points and the others is , say 5. If x_i and x_j are 2 of the other points such that x_i is distance 5 from x_{k_1} and x_j is distance 5 from x_{k_3}, then the suggested tie-breaking strategy takes $x_{k_4} = x_j$. Additionally, precautions must be taken against roundoff error inadvertently causing two mathematically equal distances to be computed non-equally in order to replicate our results in the special cases of Figs. 12 and 13.

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 single-linkage clustering. For example, cutting the longest connecting edges in Fig. 4 produces exactly the single linkage clusters for c = 4 in this data. Will ordering approaches related to other linkage algorithms work better? How does the pattern in ODI's such as those in Figs. 12 and 13 depend on the choice of distance measure in cases when relational data is generated from object data? To what extent can the information in an ODI be properly decoded so that the geometric structure underlying the data can be understood? Can existing image processing techniques be applied to an ODI to further enhance the cluster information contained in the relational data? Are there analogous visual displays that will help with the somewhat similar problem of cluster validity? Finally, how well does VAT scale up to data sets with large (but very commonly encountered) values of s and n? Most importantly, will this work lead to a killer vacation for the authors and their families in Hawaii?

REFERENCES

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

[2] B.S. Everitt, *Graphical Techniques for Multivariate Data*. New York, NY: North Holland, 1978.

[3] J.W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

[4] W.S. Cleveland, *Visualizing Data*. Summit, NJ: Hobart Press, 1993.

[5] M. Kendall and J.D. Gibbons, Rank Correlation Methods. New York, NY: Oxford University Press, 1990.

[6] I. Borg and J. Lingoes, Multidimensional Similarity Structure Analysis. New York, NY: Springer-Verlag, 1987.

[7] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," in *Journal of the Royal Statistical Society*, vol. B39, pp. 1-38, 1977.

[8] J.K. Dixon, "Pattern recognition with partly missing data," in *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 617-621, 1979.

[9] R.J. Hathaway and J.C. Bezdek, "Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm," unpublished.

[10] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*. New York, NY: Wiley, 1987.

[11] J.L. Schafer, *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.

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

[13] R.F. Ling, "A computer generated aid for cluster analysis," in Communications of the ACM, vol. 16, 355-361, 1973.

[14] J.C. Bezdek and N.K. Pal, "Some new indexes of cluster validity," in *IEEE Transactions on Systems, Man and Cybernetics-Part B*, vol. 28, pp. 301-315, 1998.

[15] J.C. Bezdek, J.M. Keller, R. Krishnapuram, L.I. Kuncheva and N.R. Pal, "Will the real IRIS data please stand up?" in *IEEE Transactions on Fuzzy Systems*, vol. 7, pp. 368-369, 1999.

[16] K.C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," in *Pattern Recognition*, vol. 24, pp. 567-578, 1991.

[17] GENLAB is at: www-it.et.tudelft.nl/~imds/index.html

[18] R. A. Johnson and D.A. Wichern, *Applied Multivariate Statistical Analysis*, 3rd Ed., Prentice-Hall, Englewood Cliffs, NJ, 1992.

[19] J. C. Bezdek, Numerical Taxonomy with Fuzzy Sets," J. Math. Biology, 1, 57-71, 1974.