Pergamon

Pattern Recognition, Vol. 28, No. 3, pp. 381-391, 1995
Elsevier Science Ltd
Copyright © 1995 Pattern Recognition Society
Printed in Great Britain. All rights reserved
0031-3203/95 $9.50+.00

0031-3203(94)00111-1

AN INDEX OF TOPOLOGICAL PRESERVATION FOR FEATURE EXTRACTION

JAMES C. BEZDEK†‡ and NIKHIL R. PAL§

† Department of Computer Science, The University of West Florida, Pensacola, FL 32514, U.S.A.
§ Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta-700035, India

(*Received* 25 *October* 1993; *in revised form* 28 *July* 1994; *received for publication* 26 *August* 1994)

Abstract—This paper is about the ability of principal components analysis, the Sammon algorithm, and an extension of the Kohonen self-organizing feature map to preserve spatial order during feature extraction on unlabeled data. Transformations to q-space that preserve the *order* of all pairwise distances in any set of vectors in p-space are defined as *metric topology preserving* (MTP) transformations. We give a necessary and sufficient condition for this new property in terms of the Spearman rank correlation coefficient. Unlike many other measures of extracted feature quality, the MTP index is independent of the extraction method. A modification of the Kohonen self-organizing feature map algorithm that extracts vectors in q-space from data in p-space is developed. The extent to which principal components, Sammon's algorithm and our extension of the *self-organizing feature map* (SOFM) preserve the MTP property is discussed. Our MTP index shows that the first two methods preserve distance ranks on seven data sets much more effectively than extended SOFM.

Feature extraction Principal components analysis Sammon's method Self-organizing feature maps Topological preservation

1. INTRODUCTION

Object data are represented as $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$, a set of ($n$) feature vectors in *feature space* $\mathscr{R}^p$. The jth observed object (some physical entity such as a fish, pickup truck, medical patient, stock market report, etc.) has vector $\mathbf{x}_j$ as it's numerical representation; x_{jk} is the kth characteristic (or *feature*) associated with object j. To characterize feature extraction, let P ($\mathscr{R}^p$) and P ($\mathscr{R}^q$) be the sets of all subsets of $\mathscr{R}^p$ and $\mathscr{R}^q$, respectively. Let Φ: $P(\mathscr{R}^p) \mapsto P(\mathscr{R}^q)$ be a set-to-set transformation with image $Y = \Phi[X] \in P(\mathscr{R}^q)$. The dimension q can be greater than, equal to, or less than p. When $|X| = |Y| = n$, there is a correspondence $\mathbf{x}_i \overset{\Phi}{\leftrightarrow} \mathbf{y}_i$ $\forall i$, and we call Φ a *feature extraction* transformation.(1-3) Usually, Φ is a point to point transformation, i.e. $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} = \{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_n)\}$, which is a special case of the general formulation. The word transformation includes two realizations: Φ may be a *function*, written as $\Phi = f$; or Φ may be an *algorithm*, denoted here as $\Phi = A$. Functions lend themselves to analysis of properties such as linearity, etc. Algorithms are computational transformations, and hence, their functional properties are generally difficult to verify. We avoid using the word 'map' as a synonym for transformation, as there is much confusion in the literature about whether the word is being used in its mathematical sense; or its perceptual sense (as a visual display, which is a 'map' in a much different context); or both.

Dimensionality is sometimes increased when p is small in order to enhance the utility of the original data. For example, simple images contain only one dependent variable (intensity) at each pixel in the image. Extracting features such as estimates of the gradient of the picture function in each coordinate direction and its average intensity over a window centered on each pixel increases the dimensionality of the raw (sensed) data from $p = 1$ to $q = 3$.

When p is large, feature extraction is used for two different but somewhat related problems: *dimensionality reduction* and *visual displays*. Reducing p to $q \ll p$ reduces the space and time complexity of computations that use the extracted data. Another issue is *redundancy*—are some of the original p features 'unnecessary'? Can we do just as well (with respect to a problem under consideration) with $q < p$ new features derived from the originals? For example, transformed features can work better than the original data for purposes such as classifier design.

Any feature extraction method that produces $Y = \Phi[X] \subset \mathscr{R}^q$ can be used to make visual displays by taking $q = 1$, 2, or 3 and plotting Y on a rectangular coordinate system. In this category, for example, are feature extraction functions such as the linear transformations defined by principal components matrices, and feature extraction algorithms such as the Sammon's. A large class of transformations, however, produce *only* visual displays from X (and not data sets $Y \subset \mathscr{R}$,

‡ Author to whom all correspondence should be addressed.

$\mathscr{R}^2$ or $\mathscr{R}^3$) through devices other than scatterplots. In this category are functions such as Andrews plots,[4] and algorithms such as Chernoff Faces,[5] and trees and castles.[6] This more limited class of transformations will be represented as f^D, A^D: $\mathscr{R}^p \mapsto V(\mathscr{R}^q)$, and these will be called, respectively, *feature display* functions and algorithms. The nature of the image space $V(\mathscr{R}^q)$ of display transformations depends on the function or algorithm being used.

For $q = 2, V(\mathscr{R}^2)$ is often a *viewing plane* which must have a coordinate system to enable visual displays, but there are not extracted vectors $\mathbf{y}_i = \Phi(\mathbf{x}_i)$ that have coordinates in $V(\mathscr{R}^2)$. For example, Andrews plots are made by representing each $\mathbf{x}_j$ in X as a trigonometric polynomial, and the visual display of X is the collection of plots of the n polynomials produced by applying f^D_{Andrews} to X. An example of a hybrid technique that combines clustering with feature extraction uses fuzzy c-means to first color (label) $X \subset \mathscr{R}^p$.[7] Then cores in X are found via membership value thresholding. And finally, any feature extraction transformation is applied to the cores to produce visual displays as core-zone scatterplots of the cluster substructure found in X.

The choice of Φ for extraction or display is often dictated by a desire to 'preserve some property' of X. For example, principal components[8,9] maximizes the preservation of sample variance; the Sammon algorithm[10,11] attempts to preserve interpoint distance pairs; and Kohonen's self-organizing feature map (SOFM[12]) is *said to* preserve a certain topological relationship amongst the data points. A derivative question *a posteriori* to extracting $\Phi[X]$ is: how well was the property of interest preserved? The goal of this work to answer this question with respect to the preservation of metric topology for these three methods.

2. THE THREE TECHNIQUES

Principal components[8,9] is arguably the most popular method in the functional category for extraction and display. This approach produces new features as linear combinations of the originals, so it is a feature extraction (and hence, feature display) function. To describe the method, let S be the sample covariance matrix of X, $S = 1/n\,\Sigma^n_{k=1}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T$, where $\mathbf{m} = 1/n\,\Sigma^n_{k=1}\mathbf{x}_k$ is the sample mean of X. Assuming S to be positive definite, extract and order the p eigenvalues of S, as, say, $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0$, and let $\mathbf{v}_i$, $i = 1, 2, \ldots, p$ be the corresponding unit eigenvectors; i.e. $S\mathbf{v}_i = \lambda_i \mathbf{v}_i$, $i = 1, 2, \ldots, p$ and $\mathbf{v}_i^T\mathbf{v}_i = 1\,\forall i = 1, 2, \ldots, p$. The eigenvectors of S are used to define p linear feature extraction functions. Let P_k denote the $p \times k$ matrix whose columns are the first k (ordered) eigenvectors of S, and define, for $q = 1, 2, \ldots, p$, $f_{\text{PC},q}$: $\mathscr{R}^p \to \mathscr{R}^q$ as $\mathbf{y}_q = f_{\text{PC},q}(\mathbf{x}) = P_q^T\mathbf{x}$. $\mathbf{y}_q$ is called the qth order principal component of $\mathbf{x}$, and $y_{qi} = \mathbf{v}_i^T\mathbf{x}$ is called the ith score or loading of $\mathbf{x}$.

For example, $f_{\text{PC},2}(\mathbf{x})$ is a point in $\mathscr{R}^2$, the set $Y_{\text{PC},2} = f_{\text{PC},2}[X]$ is a set of two-dimensional feature vectors extracted from X, and a plot of $Y_{\text{PC},2}$ is a scatterplot of the first and second principal components of X. This plot is the unique two-dimensional projection of X onto a plane through the origin that accounts for the maximum possible fraction of the total sample variance in X, namely $(\lambda_1 + \lambda_2)/\Sigma^p_{i=1}\lambda_i$. Of course *any* of the $p(p-1)/2$ pairs of PCs of the data may be plotted; of these, our notation accounts only for the single pair produced by $f_{\text{PC},2}$. For example, a scatterplot of the last two PCs is sometimes examined to see what is in the 'tail' of the sample variance. And for any $q \leq p$, $Y_{\text{PC},q} = f_{\text{PC},q}[X]$ is data in $\mathscr{R}^q$ extracted from X. Finally, observe that $f_{\text{PC},q}$ is linear, $f_{\text{PC},q}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha f_{\text{PC},q}(\mathbf{x}) + \beta f_{\text{PC},q}(\mathbf{y})$.

The Sammon method[10,11] is a set-to-set algorithmic transformation denoted by $A_{\text{S},q}$: $P(\mathscr{R}^p) \mapsto P(\mathscr{R}^q)$. $A_{\text{S},q}$ looks for vectors $Y_{\text{S},q} = A_{\text{S},q}[X]$ in $\mathscr{R}^q$ that have the same pairwise distances as their pre-images in X. Let d^*_{ij} be the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathscr{R}^p$ and d_{ij} be the distance between the corresponding (unknown) vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ produced by $A_{\text{S},q}$ in $\mathscr{R}^q$. Sammon suggested minimizing $E_{\text{S},q}(Y) = (1/\Sigma_{i<j}d^*_{if})\,\Sigma_{i<j}(d^*_{ij} - d_{ij})^2/d^*_{ij}$. The Sammon error function $E_{\text{S},q}(Y) = 0$ if and only if $A_{\text{S},q}$ preserves all $n(n-1)/2$ distances *exactly*. Thus, $A_{\text{S},q}$ attempts to be an *isometric* connector between X and $A_{\text{S},q}[X]$. The Sammon algorithm is the method of steepest descent applied to $E_{\text{S},q}(Y)$.

The Sammon algorithm is well known for its ability to find good lower dimensional representations of X, but it has not found extensive use for data sets having large values of n because each iteration attempts to solve (qn) simultaneous, coupled, non-linear equations in the unknowns $\{y_{ij}\}$. Several modifications of the Sammon algorithm that attempt to reduce its complexity have been proposed.[13-16] $A_{\text{S},q}$ provides a nice benchmark for other extraction and display algorithms, because the property it tries to capture—isometry—is probably the strongest one that might be proposed in the context of feature extraction for pattern recognition.

The SOFM and a special case of it called *Learning Vector Quantization* (LVQ) are computational networks that 'learn' vector quantizers (or class prototypes). Advantages, disadvantages and variants of LVQ have been discussed, and modified learning vector quantization algorithms have been suggested elsewhere.[17-21] This article concentrates on the topological preservation aspect of the SOFM.

SOFM is an algorithmic display method denoted here by A^D_{SOFM}: $\mathscr{R}^p \mapsto V(\mathscr{R}^q)$ that is often advocated for visualization of 'topological relationships' and distributional density in X. In principle X can be transformed onto a display lattice $O_q \subset V(\mathscr{R}^q)$ for any q; in practice, visual displays can be made only for $q \leq 3$, and are usually made on a planar configuration arranged as a rectangular or hexagonal lattice. In this article we concentrate on square $(m \times m)$ displays on the two-dimensional lattice $O_2 \subset V(\mathscr{R}^2)$. SOFM is implemented for this case through the network architecture shown in Fig. 1.

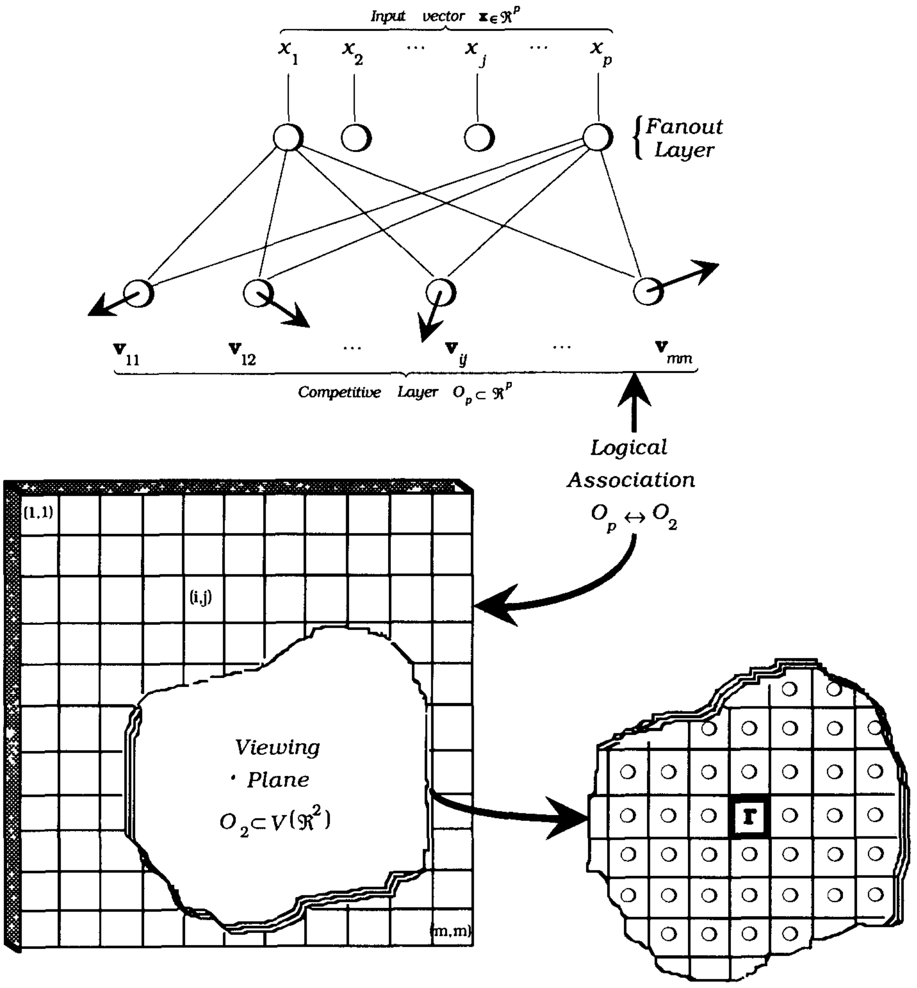

Input vector $x \in \Re^p$
x_1 x_2 ... x_j ... x_p
Fanout Layer
v_{11} v_{12} ... v_{ij} ... v_{mm}
Competitive Layer $O_p \subset \Re^p$
Logical Association $O_p \leftrightarrow O_2$
(1,1)
(i,j)
(m,m)
Viewing Plane $O_2 \subset V(\Re^2)$
r

Fig. 1. The SOFM architecture.

Input vectors $\mathbf{x} \in \mathscr{R}^p$ are distributed by a fan-out layer to each of the m^2 display nodes in the *competitive layer*. Each node in this layer has a *weight vector* or *prototype* $\mathbf{v}_{ij}$ associated with it as shown in Fig. 1. We let $O_p = \{\mathbf{v}_{ij}\} \subset \mathscr{R}^p$ denote the set of m^2 weight vectors. O_p is (logically) connected to a square display grid $O_2 \subset V(\mathscr{R}^2)$. (i,j) in the index set $\{1, 2, \ldots, m\} \times \{1, 2, \ldots, m\}$ plays two roles; it is the logical address of the cell, and it is also a geometric vector with coordinates (i,j) which we take to be the *center* of the cell (i,j). This gives a one-to-one correspondence between the m^2 p-vectors $\{\mathbf{v}_{ij}\}$ and the m^2 cells $\{(i,j)\}$, i.e. $O_p \leftrightarrow O_2$. In the literature display cells are sometimes called nodes, or even neurons, in deference to possible biological analogs.

SOFM begins with a (usually) random initialization of the weight vectors $\{\mathbf{v}_{ij}\}$. To simplify notation we suppress the double subscripts associated with $O_p \leftrightarrow O_2$. Let $\mathbf{x} \in R^p$ enter the network and let t denote the current iterate number. Find $\mathbf{v}_{r,t-1}$, the vector in O_p that best matches $\mathbf{x}$ in the sense of minimum Euclidean distance in $\mathscr{R}^p$; i.e. r is the index of the 'winner' prototype, $r = \underbrace{\arg\min}_{i} \{ \| \mathbf{x} - \mathbf{v}_{i,t-1} \| \}$. $\mathbf{v}_{r,t-1}$ has a (logical) image which is the cell in O_2 with subscript $r, t-1$. Next, a topological (spatial) neighborhood $N_t(r)$ centered at r is defined in O_2, and its display cell neighbors are located. Finally, $\mathbf{v}_{r,t-1}$ and the other weight vectors associated with cells in the spatial neighborhood $N_t(r)$ are updated using the rule $\mathbf{v}_{i,t} = \mathbf{v}_{i,t-1} + h_{ri}(t)(\mathbf{x} - \mathbf{v}_{i,t-1})$. The function $h_{ri}(t)$ expresses the strength of interaction between cells r and i in O_2. Usually $h_{ri}(t)$ decreases with t, and for fixed t it decreases as the distance (in O_2) from cell i to cell r increases. A common choice for $h_{ri}(t)$ is $h_{rt}(t) = \alpha_t e^{-\mathrm{dist}^2(r,i)\sigma_t^2}$, where α_t and σ_t decrease with time t. The extent of the topological neighborhood $N_t(r)$ also decreases with time. At the conclusion of training a final pass is made through X to get the display in $V(\mathscr{R}^2)$, which is produced by 'lighting up' (marking) each cell r in O_2 that corresponds to a winner node $\mathbf{v}_r \in O_p$.

Given a lattice structure O_2, there are different ways to define topological neighbors. For example, the cut-out in Fig. 1 depicts square neighborhoods about r. Note that the preimages of $N_t(r)$ are not necessarily metrical neighbors in R^p. However, this scheme often preserves 'spatial order' in the sense that weight vectors which are metrically close in $\mathcal{R}^p$ generally have, at termination of the learning procedure, visually close images in the viewing plane O_2. One objective of the present work is to clarify the notion of spatial order by giving a precise definition and test for it.

Convergence of SOFM has been widely studied; see reference (19) for a representative treatment. There are many variations of the basic SOFM algorithm. For example, in reference (20) an algorithm is suggested that uses metrically defined neighborhoods of winners $\mathbf{v}_r$ *in feature space* $\mathcal{R}^p$. In reference (21) neighborhoods of winner $\mathbf{v}_r$ in $\mathcal{R}^p$ are chosen from nodes in a minimal spanning tree constructed on the weight vector set $O_p = \{\mathbf{v}_{ij}\} \subset \mathcal{R}^p$. The version of SOFM that we will modify to extract feature vectors in $\mathcal{R}^q$ is:

Algorithm. A^D_{SOFM} Kohonen[12]: Display only

```
Begin
  Input $X = \{\mathbf{x}_i \in \mathcal{R}^p: i = 1, 2, \ldots, n\}$;
  Input m—the display grid size, a square m × m lattice is assumed;
  Input maxstep—maximum number of updating steps;
  Input $Nd_0$—initial neighborhood size;
  Input $\alpha_0$—the initial step size (learning coefficient);
  Input $\sigma_0$ and $\sigma_f$—parameters to control effective step size;
/** Learning phase **/
  Randomly generate initial weight vectors $\{\mathbf{v}_{ij}, i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, m\}$
  $t \leftarrow 0$;
  While (t < maxstep)
    Select randomly $\mathbf{x}(t)$ from $X$; find $r = \underbrace{\arg\min}_{i} \{\|\mathbf{x} - \mathbf{v}_i\|\}$; note that $r$ and $i$ actually stand for two
      dimensional indices that uniquely identify a weight vector in $O_p$;
    $\mathbf{v}_i(t+1) \leftarrow \mathbf{v}_i(t) + \alpha_t g_t(\mathrm{dist}(r, i))(\mathbf{x}(t) - \mathbf{v}_i(t)) \quad \forall i \in Nd_t(r)$;
    $\mathbf{v}_i(t+1) \leftarrow \mathbf{v}_i(t) \quad \forall i \notin Nd_t(r)$;
      where $dist(r, i)$ is the Euclidean distance between the geometric centers of cells $r$ and $i$ on the display
      lattice, and $g_t(d) = e^{-d^2/\sigma_t^2}$;
    $t \leftarrow t + 1$;
    $\alpha_t \leftarrow \alpha_0(1 - t/maxstep)$;
    $Nd_t \leftarrow Nd_0 - t(Nd_0 - 1)/maxstep$;
    $\sigma_t \leftarrow \sigma_0 - t(\sigma_0 - \sigma_f)/maxstep$;
/** —there are many other ways to readjust $\alpha_t$, $Nd_t$ and $\sigma_t$, and many choices for $g_t$ **/
  While End
/** Display phase **/
  For each $\mathbf{x} \in X$ find $r = \underbrace{\arg\min}_{i} \{\|\mathbf{x} - \mathbf{v}_i\|\}$, and light up (mark) cell $r$ in $O_2$.
End
```

Kohonen[12] states that there are two opposing tendencies at work in the self-organizing process. First, the weight vectors tend to describe the density function of the input population by assuming the shape of the data. And second, the local interaction between processing units in O_2 tends to preserve continuity in the double (two-dimensional) sequences of weight vectors. In other words the weight vectors in O_p are trying simultaneously to approximate the distribution or spatial ordering of the data in $\mathcal{R}^p$; and to have logical images which are topologically ordered in O_2.

3. FEATURE EXTRACTION WITH SOFM

We believe three things are needed to make the idea of topological preservation precise:

(i) an image of X under the transformation;
(ii) a definition of topological preservation; and
(iii) a way to measure the extent to which a given transformation preserves topology in the defined sense.

A^D_{SOFM} yields a display in $V(\mathcal{R}^2)$, but does *not* extract a lower dimensional representation of X. Consequently, the extent to which SOFM preserves topology cannot be measured until it is extended to extract $Y \subset \mathcal{R}^q$. A natural strategy for doing this follows (we give the proposed method for $q = 2$; it can be done for any q).

After SOFM terminates, find the winner cell in O_2 for each $\mathbf{x} \in X$ and take the coordinates of the center of that display cell as a two dimensional representation of $\mathbf{x}$. What if more than one input vector marks the same display cell? Input vectors that mark the same cell are either identical, or very close geometrically, unless the number of display cells is much smaller than the total number of input data points. Let k be the number of data points that are imaged on a particular cell in O_2. Among the k points, suppose that s are identical. In this case we generate $(k - s + 1)$ distinct two-vectors distributed randomly over (the area in $\mathscr{R}^2$ covered by) the cell. Identical points in $\mathscr{R}^p$ are given identical coordinates in $\mathscr{R}^2$.

We still need to associate each input vector with a specific point in its marked cell. One way to do this is to order the k points in $\mathscr{R}^p$ associated with the cell based on their distances from the origin, and do the same for the generated points in $\mathscr{R}^2$. Then the image vector (in O_2) with rank i for this cell is associated with the data vector having the same rank. This does not guarantee that topological relationships between the k points in $\mathscr{R}^p$ will be in any sense preserved in their two-dimensional counterparts. However, it will approximate the situation in $\mathscr{R}^p$ because not many points will be associated with the same cell, and those that are will be close to each other. One problem with this method is that the data extracted depend on algorithmic choices. Different initializations and parameters of the learning process may generate completely different extractions having very similar levels of topological preservation (for example, rotating the display lattice O_2 generates a different data set with the same neighborhood structure). We formalize this:

Algorithm $A_{\mathrm{E,SOFM}}$: extraction and display

```
Begin
  Call $A^{D}_{\mathrm{SOFM}}$; /*** Extraction Phase ****/
  for $k = 1, 2, \ldots, n$
    $r \leftarrow \underbrace{\arg\min}_{i}\{\|\mathbf{x}_{\hat{k}} - \mathbf{v}_i\|\}$
    $count[r] \leftarrow count[r] + 1$; /** count is an integer array **/
    $List(r) \leftarrow List(r)|k$; /**| stands for concatenation, i.e. append $k$ to the $r$th list **/
  for $i = 1, 2, \ldots, m^2$
    if $count[i] = 1$
      Take the centroid of the $i$-th cell as the generated vector $\mathbf{y}$ for the $\mathbf{x}$ in $List(i)$;
      /** For a two-dimensional display lattice, $i$ corresponds to a pair of coordinates, say, $(t, r)$, **/
      /** and take $\mathbf{y} = \begin{pmatrix} t-.5 \\ r-.5 \end{pmatrix}$.   **/
      else
    if $count[i] > 1$
      Using the indices stored in $List(i)$, find the number (s) of distinct $\mathbf{x}$'s for which node $i$ is the winner;
      Generate $Y = \mathbf{y}_j$, $j = 1, \ldots, s$ points randomly from the unit square around point $i$;
      /** For a two-dimensional display lattice, $i$ corresponds to a pair of coordinates, say, $(t, r)$, **/
      /** and we draw points from the box defined by $\{(t-1, r-1), (t, r)\}$ **/
      Generate $l_i$, $i = 1, 2, \ldots, s$, the list of indices of the points in $Y$ after sorting their distances from the origin;
      Generate $L_i$, $i = 1, 2, \ldots, s$, the list of indices of the distinct $\mathbf{x}$'s corresponding to $List(i)$ after sorting their
        distances from the origin;
      for $j = 1, 2, \ldots, s$
        assign $\mathbf{y}_{l_j}$ to $\mathbf{x}_{L_j}$;
End
```

Let $Y_{\mathrm{E,SOFM},q} = A_{\mathrm{E,SOFM},q}[X]$ represent the q-dimensional feature vectors extracted from X by $A_{\mathrm{E,SOFM},q}$. How good is the extracted data? The answer to this, of course, depends on what Y will be used for. In pattern recognition (clustering and classification), distance relationships are often important; now we turn to a new measure of how well distance order is preserved by each of the three methods just described.

4. TOPOLOGY PRESERVATION UNDER FEATURE EXTRACTION

Topological spaces without metrics are usually regarded as topologically *equivalent* when neighborhood structures can be reproduced with continuous transformations. When dealing with metric spaces, topological equivalence usually connotes isometry (all distances in every neighborhood are preserved). Most pattern recognition algorithms for clustering and classification depend on the notion of *distance*, so the Sammon attempt to produce isometric images of high dimensional data has the underlying rationale that every result obtained using distance-based techniques in X can be exactly replicated in any isometric image of it. However, it is very difficult to extract isometric images of data. In this section we introduce a new property that finite data sets may possess that falls in-between continuity (which may not preserve distance order) and isometry (which preserves not only order, but actual distances). The basic idea of this new property is that it requires the preservation of distance *order*-stronger than continuity, but weaker than isometry.

Much has been written about the SOFM's ability to preserve topology.[17, 18] However, a formal definition of this idea is hard to find, and harder to make, because A^D_{SOFM} does not produce feature vectors in $\mathscr{R}^q$. Properties of X are—at best—mirrored indirectly by the prototype vectors in O_p (which exhibit spatial order) or cell markings in O_2 (which exhibit logical order). We are unaware of any definition or result that relates the topological preservation ascribed to SOFM to either of the precise mathematical notions of topological equivalence just mentioned. Nonetheless, there have been several recent attempts at quantification of this idea for SOFM.[22,23]

Kraaijveld *et al.*[22] defined a way to modify the SOFM display strategy so that it produced a gray-tone image; we use $A^{D,KMJ}_{\mathrm{SOFM}}$ to represent their display algorithm. They also discussed three ways to compare the extent of topological preservation under SOFM with that of the Sammon algorithm. Two of their approaches were 'performance-based' methods using labeled data that really afford a comparison between A^D_{SOFM} and $A^{D,KMJ}_{\mathrm{SOFM}}$ at accomplishing a task such as classifier design. The connection between these two means of comparison and the notion of topological preservation is unclear to us. The third method for assessing preservation of topology reported in reference (22) used distances between the weight vectors $\{\mathbf{v}_{ij}\}$ in $\mathscr{R}^p$ to induce a measure of dissimilarity on O_2 by defining

- the distance δ_{ij} between adjacent cells i and j in O_2 as the Euclidean distance $\hat{d}_{ij}$ between their corresponding weight vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ in O_p; and
- the distance δ_{st} between two non-adjacent cells s and t in O_2 as the *minimum* of the summed distances between adjacent cells, the minimum being taken over all possible 8-connected paths in O_2 from cell s to cell t.

Once this is done, the functional form of the Sammon error measure is adapted for SOFM by computing $E_{KMJ} = (1/\Sigma_{i<j}\hat{d}_{ij})\,\Sigma_{i<j}(\hat{d}_{ij} - \delta_{ij})^2/\hat{d}_{ij}$. If we accept Sammon's error function as an index of topological preservation (certainly it is, in the strong sense of being zero only for isometric images of X under A_{S}), this allows us to (roughly) compare the extent to which A^D_{SOFM} and A_{S} preserve spatial order. We say *roughly* for two reasons. First, because A_{S} produces vectors in $\mathscr{R}^q$ that *have* neighborhoods in the same sense as X, whereas A^D_{SOFM} does not. And second, since O_2 is $m \times m$ in size, the matrix $[\delta_{ij}]$ of dissimilarities induced on it is necessarily $m \times m$. Consequently, E_{KMJ} is computable only if $\hat{D} = [\hat{d}_{ij}]$ is also $m \times m$. Generally however, $m \neq n$ [it was not in reference (22)]. Even if $\hat{D}$ comprises distances between pairs of points in X [this was not specified in reference (22)], $\hat{D}$ must be either compressed (so some of the data distances are left out) or expaneded (so some of the data distances are used more than once) to make it the right size. In consequence, E_{KMJ} is at best a very rough indirect index of topological preservation.

Another indicant of the quality of topological preservation under SOFM is reported in reference (23) by Bauer and Pawelzik, who use the topographic product as a basis for their measure. Let O_q denote a q-dimensional display lattice (conceptually, SOFM connects $O_p \leftrightarrow O_q$ for any q). Let $n_{k,O_q}(j)$ denote the index of the kth nearest neighbor (with respect to Euclidean distance on the lattice O_q) of cell j in O_q. Also let $n_{k,O_p}(j)$ be the index of the kth nearest neighbor of $\mathbf{v}_j \in O_p$. Let

$$Q_1(j,k) = \frac{d^{O_p}(\mathbf{v}_j, \mathbf{v}_{n_{k,O_q}(j)})}{d^{O_p}(\mathbf{v}_j, \mathbf{v}_{n_{k,O_p}(j)})}; \tag{1}$$

$$Q_2(j,k) = \frac{d^{O_q}(j, n_{k,O_q}(j))}{d^{O_q}(j, n_{k,O_p}(j))}; \tag{2}$$

and

$$Q(j,k) = \left(\prod_{l=1}^{k} Q_1(j,l) Q_2(j,l) \right)^{1/(2k)}, \tag{3}$$

where $d^{O_p}(\mathbf{v}_j, \mathbf{v}_{n_{k,O_q}(j)})$ is the Euclidean distance between $\mathbf{v}_j$ and the weight vector in O_p that is logically connected to the kth nearest neighbor of cell j in O_q; and $d^{O_p}(\mathbf{v}_j, \mathbf{v}_{n_{k,O_p}(j)})$ is the distance between $\mathbf{v}_j$ and its kth nearest neighbor in O_p. Further, $d^{O_q}(j, n_{k,O_q}(j))$ is the Euclidean distance on the lattice O_q between display cell j and its kth nearest neighbor in O_q; similarly, $d^{O_q}(j, n_{k,O_p}(j))$ is the distance between display cell j and the cell associated with the kth nearest neighbor of $\mathbf{v}_j$ in O_p. The topographic product of Bauer and Pawelzik for a network with m^2 weight vectors is

$$TP_{\mathrm{SOFM.BP}} \frac{1}{m^2(m^2-1)} \sum_j \sum_k Log\, Q(j,k) \tag{4}$$

$TP_{\mathrm{SOFM,BP}}$ is defined using weight vectors in O_p and cell distances in O_q. Further, the location (coordinates) of each display cell in O_q is used in (1)–(4) as a q-dimensional vector. Thus, $TP_{\mathrm{SOFM,BP}}$ measures how well the q-dimensional cell geometry in O_p matches neighborhood relationships of the weight vectors in O_p. This index is certainly not a direct function of X. Hence, we think that it does not measure the quality of preservation of any property of X after processing by SOFM.

Since the methods in references (22) and (23) do not measure topological preservation directly, we call them *indirect* indices. We seek a direct index of topological preservation for labeled or unlabeled data that is applicable to any feature extraction transformation Φ—that is, any function or algorithm that converts X in $\mathscr{R}^p$ into Y in $\mathscr{R}^q$. We begin with:

Definition MTP. Let Φ: $P(\mathscr{R}^p) \mapsto P(\mathscr{R}^q)$, $Y = \Phi[X]$, $|X| = |Y| = n > 1$. $\mathbf{x}_i \overset{\Phi}{\leftrightarrow} \mathbf{y}_i \forall i$. If d^* is a metric for $\mathscr{R}^p$ and d is a metric for $\mathscr{R}^q$, Φ is a *metric topology preserving* (MTP) transformation if and only if for any $\mathbf{x}_i$ in X, whenever $\mathbf{x}_j$ is the kth nearest (in the sense of d^*) neighbor of $\mathbf{x}_i$, then $\mathbf{y}_j$ is the kth nearest (in the sense of d) neighbor of $\mathbf{y}_i$ in Y.

When $\Phi: \mathscr{R}^p \rightarrow \mathscr{R}^q$ is a *point-to-point* transformation, a stronger statement can be made, since the vectors are not confined to belong to just X and Y. Then Φ is MTP $\Leftrightarrow \forall k$ and any $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots\ldots, \mathbf{x}_k \in \mathscr{R}^p$, $d^*(\mathbf{x}_1, \mathbf{x}) \leq d^*(\mathbf{x}_2, \mathbf{x}) \leq \ldots\ldots \leq d^*(\mathbf{x}_k, \mathbf{x}) \Rightarrow d(\Phi(\mathbf{x}_1), \Phi(\mathbf{x})) \leq d(\Phi(\mathbf{x}_2), \Phi(\mathbf{x})) \leq \ldots d(\Phi(\mathbf{x}_k), \Phi(\mathbf{x}))$. Of course, this condition also holds for set to set transfromations as long as all the points are in X and Y.

The basic idea of an MTP transformation is that it preserves in its range the *relative positions* of (all) neighbors of every point in its domain. This is a fairly strong property. An MTP transform lies in between *continuity*, which preserves neighborhoods but not distance order; and *isometry*, which preserves not only distance order, but actual distances. For example, for any $\alpha \in \mathscr{R}^+$ and $\mathbf{x} \in \mathscr{R}^p$, $\mathbf{f}(\mathbf{x}) = \alpha$ is continuous but not MTP; and $\mathbf{f}(\mathbf{x}) = \alpha\mathbf{x}$ is MTP for all $\alpha > 0$, is a contraction for $0 < \alpha < 1$, and is an isometry if and only if $\alpha = 1$.

According to our Definition, MTP transformations carry neighbors in $\mathscr{R}^p$ to neighbors in $\mathscr{R}^q$, and preserve (all) relative distance relationships. Let d^*_{ij} and d_{ij} be, respectively, the distances between $(\mathbf{x}_i, \mathbf{x}_j) \overset{\Phi}{\leftrightarrow} (\mathbf{y}_i, \mathbf{y}_j)$. If Φ is MTP, whenever d^*_{ij} is high d_{ij} should be high. The first index that comes to mind for measuring this is the statistical correlation between the $(n \times n)$ distance matrices $D^* = [d^*_{ij}]$ and $D = [d_{ij}]$. The correlation coefficient ρ for these matrices is

$$\rho(D, D^*) = \sum_{i<j} \left(\frac{(d_{ij} - \bar{d})(d^*_{ij} - \bar{d}^*)}{\sigma_d \sigma_d{}^*} \right), \tag{5}$$

where

$$\bar{d} = \frac{\sum\limits_{i<j} d_{ij}}{n(n-1)/2}, \quad \bar{d}^* = \frac{\sum\limits_{i<j} d^*_{ij}}{n(n-1)/2}, \quad \sigma^2_d = \frac{\sum\limits_{i<j} (d_{ij} - \bar{d})^2}{n(n-1)/2}$$

and

$$\sigma^2_{d^*} = \frac{\sum\limits_{i<j} (d^*_{ij} - \bar{d}^*)^2}{n(n-1)/2}.$$

Although distance order is preserved by any MTP transform, distances in Y will not normally be *uniformly* scaled with respect to distances in X. We clarify this point. Suppose $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ in $\mathscr{R}^p$ has images $Y = \Phi_1[X] = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5)$ and $Z = \Phi_2[X] = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5\}$ in $\mathscr{R}^2$ under Φ_1 and Φ_2, respectively. Y and Z might be, for example, algorithmic images under the Sammon A_S after termination at different configurations due to, say, different initializations.

This could result in the situation shown in Fig. 2(a) and (b), so Φ_1 and Φ_2 should receive exactly the same MTP score. However, these two arrangements will yield different correlation coefficients. Consequently, $\rho(D, D^*)$ is not our choice for assessing whether a feature extraction transform preserves topology in the sense of Definition MTP.

Since preservation of distance order is the important characteristic of an MTP transformation, we relabel the distances in $D^* = [d^*_{ij}]$ as $[d^*_k]$ where $d^*_k = d^*_{ij}$, $k = (i-1)((2n-i)/2) + (j-i)$. Relabel the distances in $D = [d_{ij}]$ the same way. Let the rank of the kth elements in D and D^* be $r(k)$ and $r^*(k)$, respectively, and let $\mathbf{r}$ and $\mathbf{r}^*$ be the corresponding vectors of ranks in $\mathscr{R}^{n(n-1)/2}$.

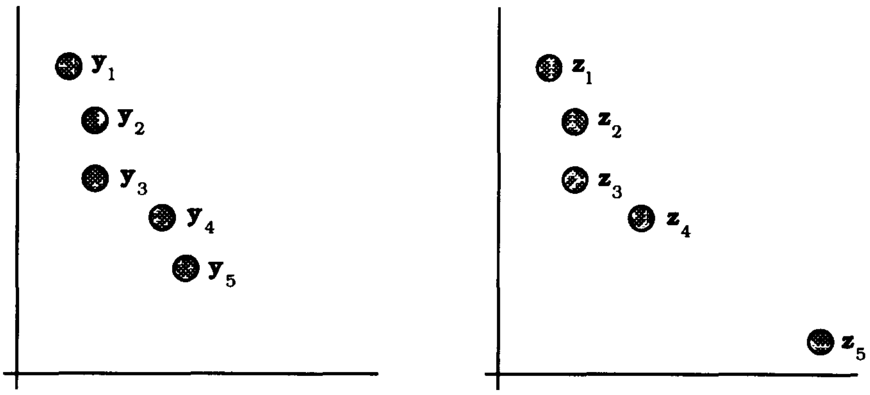
Suppose the ith nearest neighbor of $\mathbf{x}_j \in X$ is $\mathbf{x}_l$. According to Definition MTP, Φ is an MTP feature map only if, for all i and j, $\mathbf{y}_l$ is the ith nearest neighbor of $\mathbf{y}_j$. Deviation from this can be measured by a rank correlation coefficient. Specifically, either Kendall's tau or Spearman's rho[24] seems applicable. For concreteness, we choose the Spearman coefficient:

$$\rho_{\mathrm{Sp}}(\mathbf{r}^*, \mathbf{r}) = 1 - \frac{6 \sum\limits_{k=1}^{T} (r^*(k) - r(k))^2}{T^3 - T}, \tag{6}$$

where

$$T = n(n-1)/2.$$

As usual, $-1 \leq \rho_{\mathrm{Sp}} \leq 1$.[24] Here is the relationship of ρ_{Sp} to MTP transforms:



2(a) $Y = \Phi_1[X] = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\}$ **2(b)** $Z = \Phi_2[X] = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5\}$

Fig. 2. Two data sets with the same distance order structure that yield different correlation coefficients.

Theorem MTP. Let $\Phi: P(R^p) \to P(R^q)$, $\mathbf{x}_i \overset{\Phi}{\leftrightarrow} \mathbf{y}_i \forall i$. Let d^* be a metric for $\mathcal{R}^p$, $X \subset \mathcal{R}^p$ with distance matrix $D^* = [d^*_{ij}]$, and d be a metric for $\mathcal{R}^q$, $Y = \Phi[X] \subset \mathcal{R}^q$ with distance matrix $D = [d_{ij}]$, $|X| = |Y| = \mathrm{n} > 1$, and let $\mathbf{r}^*$ and $\mathbf{r}$ be the corresponding distance rank vectors in $\mathcal{R}^{n(n-1)/2}$. Φ is a *metric topology preserving* (MTP) transformation if and only if $\rho_{\mathrm{Sp}}(\mathbf{r}, \mathbf{r}^*) = 1$.

Proof. If Φ is an MTP map, then $r(k) = r^*(k)\ \forall k = 1, 2, \ldots, T$, resulting in $\rho_{\mathrm{Sp}}(\mathbf{r}, \mathbf{r}^*) = 1$. Now suppose $\rho_{\mathrm{Sp}}(\mathbf{r}, \mathbf{r}^*) = 1 \Rightarrow \left(6 \sum_{k=1}^{T} (r(k) - r^*(k))^2\right) \Big/ (T^3 - T) = 0$ $\Rightarrow \sum_{k=1}^{T} (r(k) - r^*(k))^2 = 0 \Rightarrow (r(k) - r^*(k))^2 = 0\ \forall k = 1, \ldots, T \Rightarrow r(k) = r^*(k)\ \forall k = 1, \ldots, T \Rightarrow \mathbf{r} = \mathbf{r}^* \Rightarrow \Phi$ is MTP. □

If there are ties in rank in the domain, MTP transformations will preserve them in the range. Ties in computations for ρ_{Sp} must be dealt with carefully because roundoff errors can eliminate them in the domain or range of Φ. Kendall and Gibbons[24] recommend that tied ranks be replaced by their average. This is the procedure we use in our numerical examples. As ρ_{Sp} decreases from 1, the extent to which Φ is not MTP increases until $\rho_{\mathrm{Sp}} = -1$, at which there is complete rank *reversal* between $\mathbf{r}^*$ and $\mathbf{r}$. This corresponds to 'inversion' of the order relations that MTP transformations preserve—one might call Φ such that $\rho_{\mathrm{S}} = -1$ an *anti-MTP transform.*

5. LABELED Vs UNLABELED DATA

Quantitative evaluation of extracted features for labeled data sets is easier than for unlabeled data. There are several methods available when the data are labeled.[25] For example, the natural way to assess the quality of features extracted for classifier design is to compare the performance of a classifier trained on both the original and extracted features. Machine recognition of class labels often relies on two assumption: (i) each pattern class has a homogeneous and compact shape in the feature space, and (ii) different pattern classes occupy different positions in the feature space. For labeled data, feature quality indices can be based on distances, entropies, or estimates of the within class vaiance and between class separation of the extracted features.[1,25]

As an illustration, here is a *feature evaluation index* (FEI) based on the idea that good features will simultaneously possess high *interclass* separation and low *intraclass* separation. To derive an index that measures this, let $\mathbf{v} \in \mathcal{R}^q$ be the grand mean of $Y = \Phi[X]$, and define the between scatter and within scatter matrices of Y as

$$S_B(Y) = \sum_{i=1}^{c} p_i (\mathbf{v}_i - \mathbf{v})(\mathbf{v}_i - \mathbf{v})^T; \tag{7}$$

$$S_W(Y) = \sum_{i=1}^{c} p_i \sum_{k=1}^{n_i} \frac{(\mathbf{y}_{ik} - \mathbf{v}_i)(\mathbf{y}_{ik} - \mathbf{v}_i)^T}{n_i}; \tag{8}$$

where n_i is the number points in class i, $n = \Sigma_{i=1}^{c} n_i$, $p_i = n_i/n$, $\mathbf{y}_{ik} \in \mathcal{R}^q$ is the kth data point extracted from the ith class, $\mathbf{v}_i \in \mathcal{R}^q$ is the sample mean of the ith extracted class, and c is the number of classes in the labeled data. The FEI is defined as $\mathrm{FEI}_q(Y) = \mathrm{Trace}\ S_B(Y)/\mathrm{Trace}\ S_W(Y)$.

We emphasize that for unlabeled data this type of analysis is impossible since the class label for each point in the original and extracted data is unknown. In this investigation we confine ourselves to feature extraction for unlabeled data. For unlabeled data it is possible to use indices such as $E_S(Y)$—the Sammon index—to assess the quality of features extracted with any method. However, the utility of a single value of $E_S(Y)$—which has no upper bound—for assessing the extent to which Y is not an isometric image of X is unclear to us. An important point about ρ_{Sp} is that its range is the closed interval $[-1, 1]$. Having these bounds enables us to develop a feel for the relative quality of features extracted using different methods. If the *relative ordering* of interpoint distances are preserved (i.e. if Φ is MTP) class structures, from the point of view of performance of clustering and classification algorithms, will be preserved. Our MTP index ρ_{Sp} quantifies this, and is not dependent on particular methods that extract features (or on the data being labeled).

6. SOME NUMERICAL EXAMPLES FOR UNLABELED DATA

An example of the construction of ρ_{Sp} makes the content of Theorem MTP transparent. Let $n = 5$, so $T = (n)(n-1)/2 = 10$. As k varies from 1 to $n(n-1)/2$, the addresses in D^* and D are matched in the components of $\mathbf{r}^*$ and $\mathbf{r}$. For D^* and D as given below, we have:

$$D^* = \begin{bmatrix} 0 & 3.05 & 7.31 & 2.22 & 5.66 \\ & 0 & 4.28 & 8.54 & 1.61 \\ & & 0 & 2.78 & 3.19 \\ & & & 0 & 6.91 \\ & & & & 0 \end{bmatrix}$$

$$\downarrow$$

$$\begin{array}{rcccccccccc}
\mathbf{d}^* = & (3.05, & 7.31, & 2.22, & 5.66, & 4.28, & 8.54, & 1.61, & 2.78, & 3.19, & 6.91) \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\mathbf{r}^* = & (4, & 9, & 2, & 7, & 6, & 10, & 1, & 3, & 5, & 8) \\
\mathbf{r} = & (4, & 10, & 3, & 7, & 6, & 8, & 1, & 2, & 5, & 9) \\
 & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
\mathbf{d} = & (16.1, & 37.2, & 10.3, & 18.9, & 18.6, & 19.1, & 9.15, & 9.43, & 17.1, & 27.7)
\end{array}$$

$$\uparrow$$

$$D = \begin{bmatrix} 0 & 16.1 & 37.2 & 10.3 & 18.9 \\ & 0 & 18.6 & 19.1 & 9.15 \\ & & 0 & 9.43 & 17.1 \\ & & & 0 & 27.7 \\ & & & & 0 \end{bmatrix}$$

$$\sum_{k=1}^{T} (r^*(k) - r(k))^2 = 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 2^2 + 0^2 + 1^2 + 0^2 + 1^2 = 8.$$

Substituting this into (6) gives

$$\rho_{\mathrm{Sp}}(\mathbf{r}^*,\mathbf{r}) = 1 - \left(\frac{6(8)}{1000-10}\right) = 0.9515.$$

Theorem MTP gives a necessary and sufficient condition for MTP transformations that applies to the structure extracted from X by any feature extraction transformation for labeled or unlabeled data. In particular, ρ_{Sp} can be used to assess the relative extent to which each of $f_{\mathrm{PC},q}$, $A_{\mathrm{S},q}$ and $A_{E,\mathrm{SOFM},q}$ preserves spatial order in the MTP sense.

Table 1 lists values of ρ_{Sp} for the two-dimensional sets $Y_{\mathrm{PC},2}$, $Y_{\mathrm{S},2}$ and $Y_{E,\mathrm{SOFM},2}$ extracted from the seven data sets described in the table. Apparently $A_{\mathrm{S},2}$ and $f_{\mathrm{PC},2}$ find very good two-dimensional representations of all but one of these data sets that preserve neighborhood distance structure nearly (and equally) perfectly as measured by ρ_{Sp}. The values for ρ_{Sp} on $Y_{\mathrm{S},2}[X_i]$ and $Y_{\mathrm{PC},2}[X_i]$ are nearly identical in six cases. This does not mean that $Y_{\mathrm{S},2}[X_i] = Y_{\mathrm{PC},2}[X_i]$, nor does it mean that Sammon's algorithm or principal components will always preserve distance order structures nearly perfectly, but it does imply that both algorithms extract two-dimensional data that preserve metric topology (in the sense of Definition MTP) much better than our extension of SOFM.

As $|p-q|$ increases, $p>q$, it becomes harder and harder to find q-dimensional representations of X that preserve any property. In particular, we expect ρ_{Sp} to decrease as p increases with q fixed, regardless of the method used to find Y. This point is highlighted by the last two rows of Table 1 which present the values of ρ_{Sp} for two data sets composed of 250 points drawn randomly from two and ten-dimensional normal distributions having equivalent means and covariance. Comparing rows 6 and 7 in Table 1, you can see that there is a large decrease in the (MTP) quality of the two-dimensional representation of X using any of the three extraction methods when p jumps from 2 to 10. As expected, $E_S(Y)$ increased and $(\lambda_1+\lambda_2)/\Sigma_{i=1}^{p}\lambda_i$ decreased when going from row 6 to row 7.

7. CONCLUSIONS

$A_{E,\mathrm{SOFM},2}$ is, to our knowledge, the first attempt at using SOFM to produce vectors in $\mathscr{R}^q$ as images of data in $\mathscr{R}^p$. Since the geometric coordinates of the extracted points are constrained via logical connectivity to the lattice O_q, the resultant features cannot generally be expected to be good lower dimensional representations of the data they attempt to mimic. Our index ρ_{Sp} of metric topological preservation suggests that the Sammon method and principal components preserve metric relationships much better than $A_{E,\mathrm{SOFM},2}$ (which, we emphasize, is our method, based on SOFM) for the seven data sets studied in this note. Since all but two of the data sets in Table 1 are two- or three-dimensional, generalizations of this conclusion to large values of p should be approached with caution. We found that as p increases, the ability of principal components or the Sammon algorithm to preserve MTP structure decreases. This does not detract, however, from the ability of ρ_{Sp} to assess the relative quality of features extracted from X using different methods.

Examples given elsewhere[12] indicate that SOFM's do preserve topology in the sense that display neigh-

Table 1. Values of the Spearman rank correlation coefficient ρ_{Sp}

Data set X_i	Method	ρ_{Sp}
$n = 100$ points in $\mathscr{R}^2$ uniformly distributed along the boundary of a circle centered at $(0,0)$ with radius 5.	$A_{\mathrm{S},2}$	0.999
	$f_{\mathrm{PC},2}$	0.999
	$A_{E,\mathrm{SOFM},2}$	0.758
2–100 point subsets of $\mathscr{R}^2$, each drawn uniformly from the boundary of one of the circles $x^2+y^2=16$ or $(x-15)^2+(y-15)^2=36$	$A_{\mathrm{S},2}$	0.996
	$f_{\mathrm{PC}2}$	0.999
	$A_{E,\mathrm{SOFM},2}$	0.682
The Anderson IRIS data.[26] 150 points in $\mathscr{R}^4$, $c = 3$ IRIS strains, 50 vectors for each of the three classes.	$A_{\mathrm{S},2}$	0.996
	$f_{\mathrm{PC},2}$	0.995
	$A_{E,\mathrm{SOFM},2}$	0.684
200 points in $\mathscr{R}^3$. Two subsets of 100 points each, drawn uniformly from the surface of one of two spheres. The first sphere was centered at $(0,0,0)$ with radius 4; the second was centered at $(10,10,10)$ with radius 5.	$A_{\mathrm{S},2}$	0.996
	$f_{\mathrm{PC},2}$	0.983
	$A_{E,\mathrm{SOFM},2}$	0.716
100 points in $\mathscr{R}^3$ drawn uniformly along the three-dimensional helix $x=\cos(z), y=\sin(z), z=t/\sqrt{2}$; the points were sampled at $t=0, 1, 2, \ldots, 99$.	$A_{\mathrm{S},2}$	0.999
	$f_{\mathrm{PC},2}$	0.999
	$A_{E,\mathrm{SOFM},2}$	0.712
250 points in $\mathscr{R}^2$ drawn randomly from a bivariate normal distribution having population mean vector $\binom{3}{3}$ and covariance matrix $0.04I_2$, I_2 is the two-dimensional identity matrix	$A_{\mathrm{S},2}$	0.999
	$f_{\mathrm{PC},2}$	0.999
	$A_{E,\mathrm{SOFM},2}$	0.883
250 points in $\mathscr{R}^{10}$ drawn randomly from a 10 variate normal distribution having population mean vector $(3, 3, 3, 3, 3, 3, 3, 3, 3, 3)^{\mathrm{T}}$ and covariance matrix $0.04I_{10}$, I_{10} is the ten dimensional identity matrix.	$A_{\mathrm{S},2}$	0.661
	$f_{\mathrm{PC},2}$	0.527
	$A_{E,\mathrm{SOFM},2}$	0.203

borhoods in O_2 become organized to reflect proximity among subsets of the weight vectors $\{\mathbf{v}_{ij}\} = O_p$. To us, topological preservation has a much more precise and well defined meaning, viz., relative ordering of interpoint distances are preserved under transformations of one topological space to another. Our definition of an MTP transformation is quite general, and the Spearman rank correlation coefficient provides a convenient test and measure of this property. Consequently, this notion of topological preservation should be useful in any context involving feature extraction.

Finally, we think that MTP is an important and reasonable property from the standpoint of pattern recognition. For example, any clustering algorithm (e.g. hard/fuzzy c-means, learning vector quantization, the linkage clustering algorithms, and so on) that uses a distance-based criterion to make partitioning decisions using X in $\mathscr{R}^p$ will usually make roughly the same decisions if it is instead applied to the image Y in $\mathscr{R}^q$ of an MTP transformation.

To conclude, we emphasize once more that the main objective of this article was to introduce a well-defined property (MTP) of extracted data that, when satisfied, preserved an important property (distance order) for pattern recognition. And we have characterized the MTP property quantitatively with an index (ρ_{SP}) that has crisp lower and upper bounds. Because of the bounds, our index enables users to rank the relative efficacy of different feature extraction methods at metric topology preservation for both labeled and unlabeled data. A natural corollary to this study would be to develop an algorithm based on maximizing ρ_{SP}, just as Sammon's algorithm attempts to minimize $E_S(Y)$. This will be the topic of a future report.

REFERENCES

1. R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York (1973).
2. J. W. Tukey, *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts (1977).
3. B. S. Everitt, *Graphical Techniques for Multivariate Data,* North-Holland, New York, (1978).
4. D. F. Andrews, Plots of high dimensional data, *Biometrics* **28**, 125–136 (1972).
5. H. Chernoff, The use of faces to represent points in k-dimensional space, *J. Amer. Stat. Assoc.* **68**, 361–368 (1973).
6. B. Kleiner and J. A. Hartigan, Representing points in many dimensions by trees and castles, *J. Amer. Stat. Assoc.* **76**, 260–269 (1981).
7. J. C. Bezdek and E. Chiou, Core zone scatterplots: a new approach to feature extraction for visual displays, *Comput. Graphics Image Process.* **41**, 186–209 (1988).
8. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis,* Prentice-Hall, Englewood Cliffs, New Jersey (1988).
9. K. Fukunaga and W. L. G. Koontz, Application of the Karhunen–Loeve expansion to feature selection and ordering, *IEEE Trans. Comput.* **19**, 311–318 (1970).
10. J. W. Sammon, A nonlinear mapping for data structure analysis, *IEEE Trans. Comput.* **18**, 401–409 (1969).
11. D. H. Foley and J. W. Sammon, An optimal set of discriminant vectors, *IEEE Trans. Comput.* **24**, 271–278 (1978).
12. T. Kohonen, *Self-Organization and Associative Memory.* Springer-Verlag (1989).
13. B. Schachter, A nonlinear mapping algorithm for large data bases, *Comput, Graphics Image Process.* **7**, 271–278 (1978).
14. C. E. Pykett, Improving the efficiency of Sammon's nonlinear mapping by using clustering archetypes, *Electron. Lett.* **14**, 799–800 (1980).
15. C. L. Chang and R. C. T. Lee, A heuristic relaxation method for nonlinear mapping in cluster analysis, *IEEE Trans. Syst., Man Cybern.* **3**, 197–200 (1973).
16. G. Biswas, A. K. Jain and R. C. Dubes, Evaluation of projection algorithms, *IEEE Trans. PAMI* **3**, 701–708 (1981).
17. T. Kohonen, Self-organized formation of topologically correct feature maps, *Bio. Cybern.* **43**, 59–69 (1982).
18. T. Kohonen, The self-organized map, *Proc IEEE* **78**(9), 1464–1480 (1990).
19. H. Ritter and K. Schulten, Convergence properties of Kohonen's topology conserving maps: fluctuations, stability, and dimension selection, *Bio Cybern.* **60**, 59–71 (1988).
20. N. R. Pal, J. C. Bezdek and E. C. Tsao, Generalized Clustering networks and Kohonen's self-organizing scheme, *IEEE Trans. Neural Nets.* **4**(4), 549–558 (1993).
21. J. A. Kangas, T. Kohonen and J. T. Laaksonen, Variants of self-organizing maps, *IEEE Trans. Neural Nets.* **1**(1), 93–99 (1990).
22. M. A. Kraaijveld, J. Mao and A. K. Jain, A nonlinear projection method based on Kohonen's topology preserving maps, *Proc. ICPR* B41–B45, Holland (1992).
23. H. U. Bauer and K. R. Pawelzik, Quantifying the neighborhood preservation of self-organizing feature maps, *IEEE Trans. Neural Nets.* **3**(4), 570–579 (1992).
24. M. Kendall and J. D. Gibbons, *Rank Coorelation Methods,* Oxford University Press, New York (1990).
25. P. A. Devijver and J. Kittler, *Pattern Recognition—A Statistical Approach,* Prentice-Hall, London (1982).
26. E. Anderson, The Irises of the Gaspe Peninsula, *Bulletin of the American IRIS Society,* **59**, 2–5 (1935).

About the Author—JAMES C. BEZDEK received his Ph.D. from Cornell University in 1973. His interests include pattern recognition, fishing, computational neural networks, skiing, image processing, blues music, medical computing and motorcycles. He is founding editor of the *IEEE Trans. Fuzzy Systems,* and a fellow of the IEEE.

About the Author—NIKHIL R. PAL is an Associate Professor in the Machine Intelligence Unit of Indian Statistical Institute, Calcutta. He obtained his B.Sc. (Hons) in physics and the M.B.M. (operations research) in 1979 and 1982, respectively from the University of Calcutta. He received his M.Tech. and Ph.D. degrees in computer science from the Indian Statistical Institute, Calcutta, in 1984 and 1991, respectively. He was

with the Hindusthan Motors Ltd. W.B., from 1984 to 1985 and with the Dunlop India Ltd., W.B., from 1985 to 1987. In 1987 he joined the Computer Science Unit of Indian Statistical Institute, Calcutta. During August 1991–February 1993 he visited the University of West Florida and currently also visiting the same university. He was a guest lecturer at the University of Calcutta. His research interests include image processing, pattern recognition, fuzzy sets and systems, uncertainty measures, genetic algorithms, and neural networks. He is an associate editor of the *IEEE Transactions on Fuzzy Systems* and the *International Journal of Approximate Reasoning.*