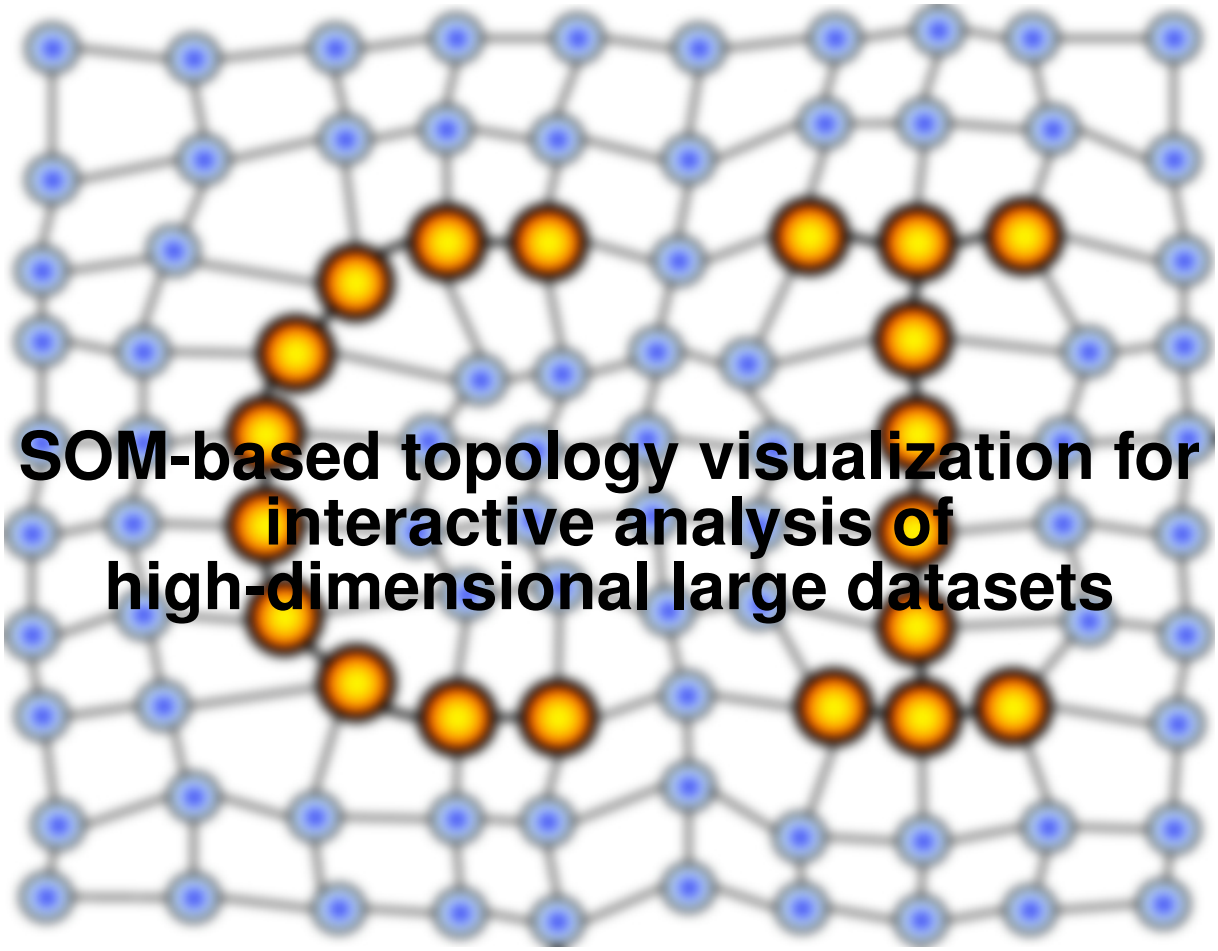
MACHINE LEARNING REPORTS



SOM-based topology visualization for interactive analysis of high-dimensional large datasets

Report 05/2012
Submitted: 24.12.2012
Published: 31.12.2012

Kadim Taşdemir[1]∗, Erzsébet Merényi[2]
(1) Department of Computer Engineering, Antalya International University
Universite Caddesi No: 2, Dosemealti, Antalya, 07190, Turkey
(2) Dept. of Statistics and Dept. of Electrical and Computer Engineering, Rice University
6100 Main St., Houston, TX, 77005, USA
∗ corresponding author, *email: kadim.tasdemir@antalya.edu.tr*

Abstract

Low-dimensional (2- or 3-dimensional) visual representations of large, high-dimensional datasets with complicated cluster structures play a fundamental role in the discovery and identification of such structures. Visualization exploits the unmatched pattern recognition capability of humans for accurate and detailed cluster extraction, which is not possible with current automated methods because the latter still lack the power of the exceptional human reasoning. For explanatory and interactive visualization, a powerful tool is the use of self-organizing maps (SOMs). In general, by producing a spatially ordered set of quantization prototypes of large, higher-dimensional data, SOMs enable the visualization of various similarity information (such as prototype distances, distribution, topology) on a rigid lattice, without reducing the feature dimensionality. Information discovery further depends on the expressive power of the similarity measure and its visual representation. In this study, we compare the capabilities of our recent SOM visualization scheme, CONNvis, with prominent dimensionality reduction methods and show its superiority for visual assessment of intricate cluster structures.

SOM-based topology visualization for interactive analysis of high-dimensional large datasets

Kadim Taşdemir[a], Erzsébet Merényi[b]

[a] *Department of Computer Engineering, Antalya International University*
Universite Caddesi No: 2, Dosemealti, Antalya, 07190, Turkey
kadim.tasdemir@antalya.edu.tr
[b] *Dept. of Statistics and Dept. of Electrical and Computer Engineering, Rice University*
6100 Main St., Houston, TX, 77005, USA

Abstract

Low-dimensional (2- or 3-dimensional) visual representations of large, high-dimensional datasets with complicated cluster structures play a fundamental role in the discovery and identification of such structures. Visualization exploits the unmatched pattern recognition capability of humans for accurate and detailed cluster extraction, which is not possible with current automated methods because the latter still lack the power of the exceptional human reasoning. For explanatory and interactive visualization, a powerful tool is the use of self-organizing maps (SOMs). In general, by producing a spatially ordered set of quantization prototypes of large, higher-dimensional data, SOMs enable the visualization of various similarity information (such as prototype distances, distribution, topology) on a rigid lattice, without reducing the feature dimensionality. Information discovery further depends on the expressive power of the similarity measure and its visual representation. In this study, we compare the capabilities of our recent SOM visualization scheme, CONNvis, with prominent dimensionality reduction methods and show its superiority for visual assessment of intricate cluster structures.

Keywords: visualization, interactive clustering, self-organizing maps, data topology, high-dimensional data, CONNvis

1. Dimensionality reduction versus volume reduction for visualization of large datasets

Low-dimensional (2 or 3-dimensional, denoted as 2D or 3D from here on) visual representations of large, high-dimensional datasets with complicated cluster structures play a fundamental role in the discovery and identification of such structures. Visualization exploits the unmatched pattern recognition capability of humans. While cluster capture with a human-in-the-loop approach is necessarily interactive, thus slower than automated methods, the insight gained by human interaction is unparalleled. In many cases detailed structure extraction is not possible with current automated clustering methods. Therefore, it is of great importance to bring to the visualization as much knowledge about the structure of the data as possible, through intelligent information extraction and representation in a low-dimensional space.

For visualization of high-dimensional data in low-dimensional representations, there has been much research on manifold learning based on dimensionality reduction, where all n-dimensional data vectors in a data set are transformed into 2D or 3D data vectors. This is motivated by the idea that the data may lie on a low-dimensional manifold embedded in a high-dimensional space. Dimensionality reduction methods such as multi-dimensional scaling (MDS) (Cox and Cox, 2001), Isomap (Tenenbaum et al., 2000), locally linear embedding (LLE) (Roweis and Soul, 2000), Hessian LLE (hLLE) (Donoho and Grimes, 2003) are mainly developed for reconstruction of a single underlying low-dimensional submanifold rather than for visual discrimination and discovery of various structures — clusters — that may exist in the data. Hence, they are often suboptimal for discovery of cluster structure, and for classification, as shown in various papers (Yang, 2002; Polito and Perona, 2001; Vlachos et al., 2002) that propose augmentations to previously published manifold learning / dimensionality reduction methods. For example, Yang (2002) and Zhang et al. (2004) additionally use Fisher linear discriminant analysis for face recognition. Vlachos et al. (2002) modified Isomap and LLE so that both local and global distances are considered for better visualization and classification. However, the performances of the modified Isomap and LLE are not very promising for identifying different patterns due to the use of the same reconstruction objective (single underlying manifold) as with the originial Isomap and LLE. Since the data clusters may lie in different submanifolds, visualization of the separation boundaries between groups of patterns is of far greater interest for structure discovery than showing the precise underlying manifold. A novel approach for visualization of datasets with different underlying manifolds is t-distributed SNE (t-SNE) (van der Maaten and Hinton, 2008), a variation of stochastic neighbor embedding (Hinton and Roweis, 2002) with an easier-to-optimize cost function using student-t distribution for similarity calculation. Despite the non-convexity of the t-SNE cost function, t-SNE representation of local similarities provides better visual separation of the data

clusters than other dimensionality reduction methods (van der Maaten and Hinton, 2008).

A major issue with all dimensionality reduction methods for interactive visualization is that they scale exponentially with the number of data points. This makes them infeasible for large datasets (such as remote sensing images, text documents, streaming video, flow simulations, analysis of computer code performance, etc.). In contrast, prototype based methods (where the data are first summarized in prototype vectors through vector quantization (VQ)) exploit the knowledge encoded in the data representatives for interactive visualization, which can have very attractive scaling properties for the visualization of large datasets, depending on the VQ approach. Adaptive VQ algorithms, which show the data topology on the prototype level and aim to faithfully represent the local similarities of the quantization prototypes, are well suited for interactive data analysis (Kohonen, 1997; Martinetz et al., 1993; Cottrell et al., 2006; Aupetit, 2006; Bishop et al., 1998). These algorithms are either inspired by nature as in the case of Self-Organizing Maps (SOMs) (Kohonen, 1997), derived as stochastic gradient descent from a cost function as in Neural Gas (Martinetz et al., 1993) and its batch version (Cottrell et al., 2006), or obtained by expectation-maximization algorithm (Bishop et al., 1998; Aupetit, 2006). Variants of these methods, which use a magnification factor in quantization, are also analyzed to enhance the representation of complex structures including rare patterns (Merényi et al., 2007b; Villmann and Claussen, 2005; Hammer et al., 2007).

SOMs (Kohonen, 1997) stand out in the representation of the data structure because they have two advantageous properties: providing an adaptive vector quantization that results in a placement of prototypes in the data space that follows the data distribution; and ordering of these prototypes on a rigid low-dimensional lattice according to their similarity relations. Due to these properties, the density distribution – and therefore the structure – of a high-dimensional manifold can be mapped and visualized on a low-dimensional grid without reducing the dimensionality of the data vectors. This allows capture of complicated cluster structures in high-dimensional space through interactive visualizations.

Various components of the learned SOM's knowledge are often processed through visualizations to capture clusters interactively. Our visualization scheme, CONNvis, first proposed in (Taşdemir and Merényi, 2009), is different from other SOM visualization schemes in that it represents the data distribution on a subprototype level, and it achieves detailed delineation of cluster boundaries by rendering the data topology on the SOM lattice. The CONNvis is successful in interactive clustering of datasets with complex structures. We show this through data sets of progressive dimensionalities and complexities (cluster structures), culminating with a 28-cluster, 8-band, large real remote sensing image; and we show and analyze comparisons with prominent dimension reduction approaches.

Section 2 briefly reviews SOMs and their interactive visualizations. Section 3 summarizes CONNvis. Section 4 presents visualizations of five different data sets by eight dimensionality reduction methods and by CONNvis, pointing out the particular challenges in each data set, and then compares computational complexities of the presented knowledge representation methods. Section 5 concludes the paper.

2. SOM for visualization of high-dimensional data

In general, dimensionality reduction does not reduce data volume, and conversely, VQ methods do not reduce feature space dimensionality. Ideally, an effective visualization of a large, complicated, high-dimensional dataset should utilize both. Self-Organizing Maps (SOMs) provide a unique combination of adaptive vector quantization of the data space and topological ordering of the quantization prototypes on a lower-dimensional grid. This enables visualization of the topology of high-dimensional data spaces without needing dimensionality reduction of the data vectors. Thus, visualization on the SOM grid does not depend on the dimensionality of the data space directly, and therefore is not limited to data reduced to 2 or 3 dimensions.

2.1. Self-Organizing Maps

The SOM algorithm can be briefly summarized as follows: Let $\mathcal{M} \subset \mathbb{R}^d$ be a d-dimensional data manifold, and $\mathcal{G}$ be the (lower-dimensional) fixed SOM lattice of N neural units. Each neural unit j has an associated weight vector w_j which is adapted through a learning process as defined by Kohonen (Kohonen, 1997). The process consists of cycling through two steps: *i)* finding the best matching unit (BMU) w_i for a randomly picked data vector $v \in \mathcal{M}$, such that

$$\|v - w_i\| \leq \|v - w_j\| \quad \forall j \in \mathcal{G} \tag{1}$$

and *ii)* updating w_i and its neighbors according to

$$w_j(t+1) = w_j(t) + \alpha(t) h_{i,j}(t)(v - w_j(t)) \tag{2}$$

where t is time, $\alpha(t)$ is a learning parameter and $h_{i,j}(t)$ is the neighborhood function, often defined by a Gaussian kernel around the best matching unit w_i. After learning, the weight vectors become the vector quantization prototypes of the data manifold $\mathcal{M}$. Fig. 1 shows an example organization of the SOM quantization prototypes in the data space and in a 2D rectangular SOM lattice.

In our work, we use the Conscience SOM (DeSieno, 1988) for real data, because it achieves undistorted density matching (with the precision afforded by the given number of prototypes). While the Kohonen SOM follows a $Q(\mathbf{w}) \sim P(\mathbf{v})^\alpha$ power law with $\alpha = 2/3$ magnification exponent (Ritter and Schulten, 1986), the Conscience SOM produces $\alpha = 1$ for higher-dimensional, complex data as demonstrated in (Merényi et al., 2007b; Merényi, 2000).

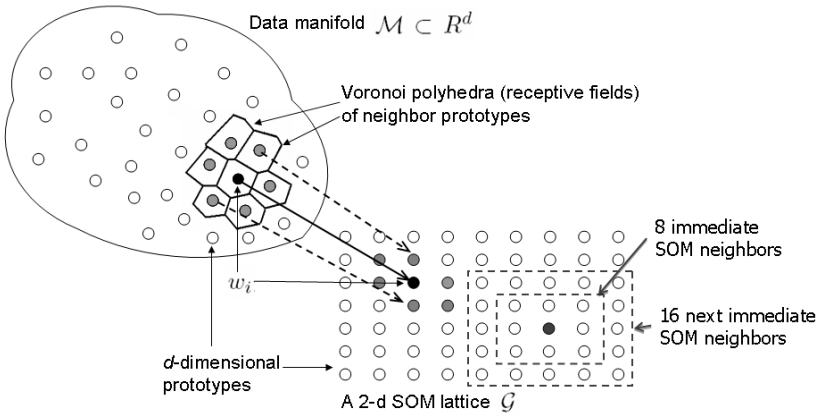


Figure 1: Organization of the SOM quantization prototypes (shown by circles) in the data space and in the SOM lattice. Each prototype is the centroid of its Voronoi polyhedron (receptive field). The prototypes that are neighbors in the data space, *i.e.*, their Voronoi polyhedra share an edge, are (ideally) neighbors in the SOM. When there are more than 8 neighbors in the data space, (with correct SOM learning) 8 of them will be the 8 immediate SOM neighbors in a rectangular lattice, and the remaining ones are expected to be in the next immediate SOM neighborhood.

2.2. Interactive Visualization Schemes for Self-Organizing Maps

The SOM facilitates the visualization of the structure of a higher-dimensional data space in lower (usually one or two) dimensions by preserving the neighborhood relations on a rigid lattice. However, post-processing is required for informative representation of the raw knowledge of the SOM. The two SOM knowledge components (most) commonly visualized are Euclidean distances of those prototypes that belong to immediate neighbor neurons in the SOM lattice; and the size of the receptive fields of the SOM neurons (data density). Various schemes have been proposed to display these quantities over the SOM lattice.

The most common, and earliest, method for displaying the Euclidean (data space) distances — or dissimilarities — of prototypes of neighboring SOM neurons is the U-matrix (Ultsch, 1993), which colors the grid cell of a prototype w_i to a gray shade proportional to the average distance of w_i to its SOM neighbor prototypes. The U-matrix and various subsequent modifications (e.g., (Kraaijveld et al., 1995; Ultsch, 2003)) work well for small data sets with a low number of clusters mapped to a relatively large SOM grid. However, they tend to obscure finer delineations between clusters in complicated and large data sets because of averaging of prototype distances over neighboring SOM grid cells, or thresholding (Merényi et al., 2007b). Another variant changes the size of the grid cells according to the average distances to neighbors (Hakkinen and Koikkalainen, 1997) but this has similar problems as the U-matrix, with less visual separation. Cottrell and de Bodt (1996) use polygons for grid cells where the distance of the vertices from the cell center are inversely proportional to the distances of each of eight immediate SOM neighbor prototypes. (The boundaries of neighboring cells are "pushed apart" — the cells are shrunken – proportional to the distances of the respective pair of SOM neighbor prototypes in each lattice direction.) This method works well for manual cluster extraction for relatively simple data sets. For identification and interpretation of clusters on the SOM lattice, an enhanced version of the U-matrix views the SOM as an undirected planar graph and clusters as connected components where a connected component is a subgraph in which two vertices (prototypes) are connected to each other by a path constructed by using the gradient of the smoothed U-matrix (Hamel and Brown, 2011). These connected components are then overlaid on the U-matrix, resulting in improvement in cluster identification especially in case of high-dimensional real world data with small number of data samples. Himberg (2000); Kaski et al. (1998, 2000); Villmann and Merényi (2001) apply automated color assignments to prototype distances for exploration of the approximate cluster structure. Some methods examine the distances based on individual feature component planes of the SOM to discover information specific to the corresponding component, which may be hidden when all planes are examined together (Kaski et al., 1998; Vesanto, 1999).

Visualization of the receptive field sizes of prototypes has been proposed in the form of data histograms by drawing vertical bars or curves, and by using gray shading in the grid cells (e.g., (Kaski et al., 1998; Vesanto, 1999; Ultsch, 2003)). For a more precise visualization of the density distribution, Pampalk et al. (2002) smooth data histograms by assigning a weighted membership of data vectors to the prototypes. Data histograms of the prototypes may conceal finer structure in complicated data since the density representation on the prototype level does not reveal prototype similarities. Therefore, data histograms are also employed together with prototype distances in the same visualization such as in Cottrell and de Bodt (1996) and Merényi et al. (2007b). Yet, there is more information, learned by the SOM, that remains untapped by the above visualizations.

In order to provide visual separation of the clusters without post-processing the SOM, Adaptive Coordinates (Merkl and Rauber, 1997) and the Double SOM (Su and Chang, 2001) update not only the prototypes but also their positions in the SOM

lattice while learning. By these methods, the SOM does not have a rigid grid anymore and the dissimilarities between the prototypes are visually exposed by their lattice distances. However, it is uncertain how these methods would work for large data volumes and for high-dimensional data. Another variant of the SOM that enables a direct and visually appealing measure of inter-point distances on the grid is the visualization induced SOM (ViSOM) (Yin, 2002). The ViSOM produces a smooth and evenly partitioned mesh through the data points which reveals the discontinuities in the manifold. The ViSOM is computationally complex for large datasets due to the requirement of a large number of prototypes even for small datasets.

SOMs have been used for more than two decades for data visualization. However, most studies described above utilize less than the full potential of the SOM knowledge. Our work, CONNvis, first proposed in Taşdemir and Merényi (2009), focuses on advancing SOM-based visualization by

- enriching the knowledge representation with resources (data topology and detailed local data distribution) not customarily used,
- employing a (density-based) similarity measure for prototypes that is not a distance based metric, and
- increasing the quantitative aspect of the evaluation of the visualization by computing "natural" thresholds (as opposed to user selected) from the data characteristics.

We summarize CONNvis next.

3. CONNvis: visualizing the connectivity structure of the data manifold

CONNvis is a graph-based similarity visualization for SOM prototypes. The graph rendered on the SOM grid represents prototype similarities, defined by the *connectivity measure* CONN. CONN is derived from the neighborhood relations (topology) of the data manifold together with the local density distribution. CONNvis enables evaluation of similarities between prototypes that are not SOM grid neighbors but are neighbors in the data manifold. It also shows a ranking of prototype similarities, and separation between submanifolds (through lack of connection between prototypes).

The connectivity matrix (Taşdemir and Merényi, 2009), CONN, is a refinement of the *induced* Delaunay triangulation of the prototype vectors, which was defined in (Martinetz and Schulten, 1994) as the intersection of the Delaunay triangulation with the data manifold. CONN assigns weights to the edges of the induced Delaunay graph. The weight of an edge connecting two prototypes is the number of data samples for which these two prototypes constitute a pair of best-matching unit (BMU) and second BMU. CONN is the matrix representation of this weighted Delaunay graph where each element, $CONN(i, j)$, is the *connectivity strength* between prototypes w_i , w_j and is equal to the weight of the edge between i and j. Formally,

$$CONN = |RF_{ij}| + |RF_{ji}| \quad (3)$$

where RF_{ij} is that section of the receptive field (Voronoi polyhedron) of the prototype w_i where w_j is the second BMU, and $|RF_{ij}|$ is the number of data vectors in RF_{ij}. CONN thus *i)* shows the data structure as expressed by the Delaunay graph of the SOM prototypes (whose distribution follows the data density), *ii)* indicates local connectivities of the manifold by expressing how the data is distributed within the receptive fields with respect to prototypes that are neighbors in the data space. (Neighbor prototypes in the data space are those which are centroids of adjacent Voronoi cells. See Martinetz and Schulten (1994) for exact definition). It facilitates identification of the discontinuities within the data set which indicate natural partitions in the data. This is in contrast to the density representations mentioned above, which express distribution on the prototype level, thus giving neither information of local density anisotropies, nor any sense of topology violations.

For visualization, CONN is rendered on the 2D rigid SOM lattice by connecting the grid locations of the SOM prototypes with lines of various widths and colors. The line widths are proportional to connectivity strengths, $CONN(i, j)$, therefore reflect the density distribution among the connected prototypes, showing the *global importance* of the connections. The connections of a prototype w_i are ranked according to their strengths to reveal the most-to-least similar neighbors (Voronoi neighbors in data space) to w_i. The rankings are indicated either by line colors, red, blue, green, yellow and dark to light gray levels or by dark to light gray levels. Since the ranking does not depend on the size of the receptive field of w_i, but only on the relative contribution of each neighbor, line colors indicate the *local importance* of a prototype's connections. Line widths and colors effectively represent the intricate details of the data structure, and enable relatively easy interpretation and capture by interactive visual clustering. In addition, topology preservation of the SOM mapping can be assessed using the line lengths and the grid neighborhood, as detailed in (Taşdemir and Merényi, 2009).

For low-dimensional (1D to 3D) datasets, CONN can also be visualized in the data space by connecting the locations of prototypes, similarly to its rendering on the SOM lattice. Fig. 2 shows example representations for two simple 2D datasets, and for the well-known 3D "Chainlink" dataset. These datasets, available at http://www.uni-marburg.de/fb12/datenbionik/data, are constructed by (Ultsch, 2005). The first dataset "Lsun" has three well-separated clusters, two rectangular and one spherical. The second dataset, "Wingnut", has two rectangular clusters with inhomogeneous density distribution within clusters and similar intra-cluster and inter-cluster distances. For both cases, the cluster structure can be expressed by CONN, as seen through its visualization, CONNvis, both in the data space and on the SOM, in spite of the variations across clusters (different shapes, proximities and inhomogeneous density distribution, respectively). For the "Chainlink" dataset, CONNvis indicates the separation between the two rings as well as the topological ordering of the data points on the SOM, by connecting the two ends of a ring with a prototype lying in the middle of the other ring. Line widths are best binned when the number of data

points is much larger than the number of prototypes, as in that case individual connectivity strengths cannot be discerned visually. Binning can be done with automated thresholding based on internal data characteristics, which results in each bin reflecting the global importance of one rank of connections. This binning method is described in detail in (Taşdemir and Merényi, 2009). The resolution of the selected thresholds not only distinguishes strong connections but also reveals weak connections between (separated) clusters.

An example for a more complex case than in Fig. 2 is given through a simple synthetic spectral image, which has 128×128 pixels with a 6D feature vector at each pixel. It has 20 spectral classes distributed spatially as shown in Fig. 3.a. The mean signatures of these classes, displayed to the right of the class map, are quite similar to each other, which poses a clustering challenge. 4 of the 20 classes, P, R, Q, S and T, are relatively small, class R has only one pixel, for additional complexity. Details of this data set are at http://terra.ece.rice.edu and in (Merényi et al., 2007b). A 20×20 SOM is used to obtain the quantization prototypes of this data. The statistics of the ranked connectivity strengths indicate that the maximum number of connections for a prototype is 16 (Fig. 3.b). The average connectivity strength is as high as 37 ($= \mu_1$) for the first ranking connections whereas it drops sharply after the fourth-ranking connections ($\mu_4 = 6$). The CONNvis of this data is obtained by using a 4-level binning scheme with $\mu_1, \mu_2, \mu_3, 0$ as the thresholds. Visual inspection of the CONNvis in Fig. 3.c. immediately reveals strongly connected groups of prototypes (clusters) with some weak connections across the clusters.

3.1. Interactive clustering from CONNvis

Graph-based visualization of prototype similarities (determined according to the data manifold) on the SOM grid enables their interactive evaluation to find separation boundaries between different patterns in the dataset. Interactive clustering from CONNvis is performed by removing weak connections (represented by the line widths) which are of low importance in terms of cluster similarities. First, topology violations (identified by line lengths) are investigated and weak global ones are removed. Any strong global connections (not present in the SOMs of the datasets in Figures 2–3) are examined further to determine whether they are due to defective learning or correctly express a structural relationship. For example, the CONNvis of Lsun or of Wingnut, shown in Fig. 2 have no topology violations for mapping of 2D datasets to 2D surfaces. The CONNvis of the 20-class dataset (Fig. 3) has many —but weak— violations, mostly contained within the obvious clusters. The global violations for this dataset are those connections with *length* > 2 since a prototype has at most 16 neighbors in the data space (from Fig. 3.b), and these 16 can fit in the second immediate SOM neighborhood as shown in Fig. 1. We remove the weak global connections, which (in this case) are those with $CONN(i, j) < \mu_4$, and visually determine strongly connected groups of prototypes (coarse clusters).

Remaining weak connections across coarse clusters need to be analyzed and removed for crisp delineation. The connections of a prototype w_j at the boundaries of k clusters are removed as follows and as illustrated for k=2 in Fig. 4.a:

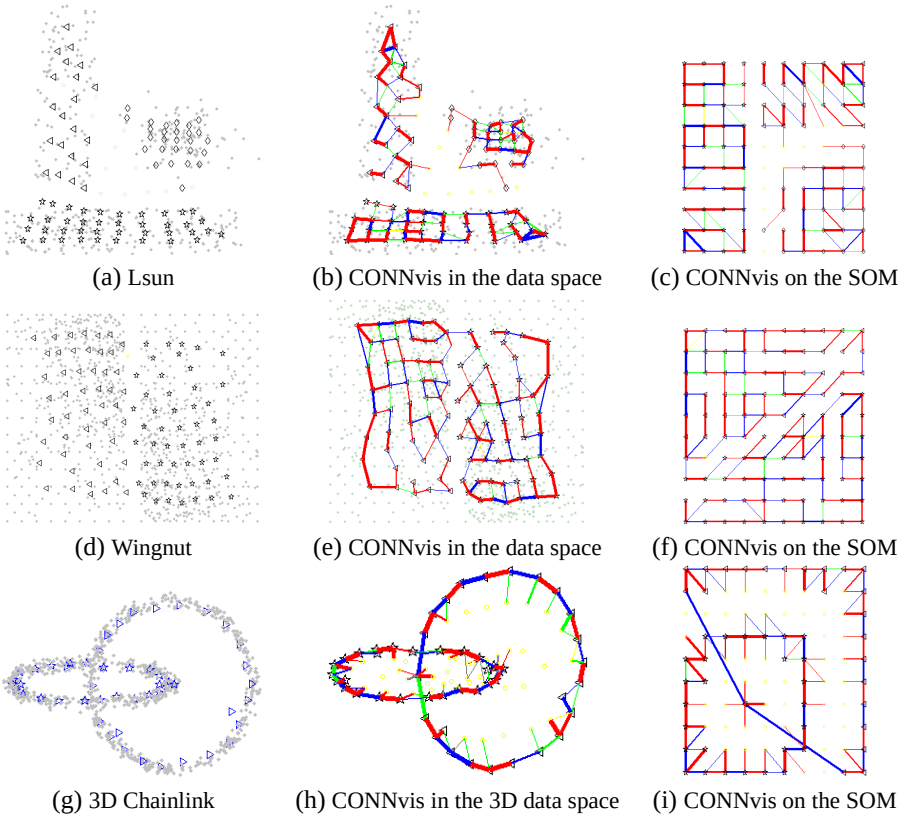
1. If the numbers of connections of a boundary prototype w_j to each of k clusters differ, keep the connections to the cluster with the highest numbers of connections, and remove all connections to other clusters.
2. If the number of connections to each cluster is the same but the connections have different strengths, keep the strongest set of connections, and remove all connections to other clusters.
3. If both the number and the strengths of connections to each cluster is the same then keep the highest ranking set of connections and remove the rest.

Fig. 4.b-c shows this interactive clustering on the 20-class dataset. The prototypes at the boundaries of coarse clusters are highlighted in black. Removal of the connections according to the above procedure results in the extraction of all 20 known clusters in the data, including the one-pixel cluster R.

4. Comparison of CONNvis with different dimensionality reduction methods for visualization in 2D space

We compare the dimensionality reduction methods tSNE, SNE, Isomap, Sammon's mapping, PCA, Kernel PCA, LLE and Hessian LLE with the quantization based topology visualization by CONNvis, for interactive visual assessment of cluster separation. To obtain the mappings produced by the dimensionality reduction methods we use the Matlab toolbox developed by van der Maaten and Hinton (2008) (available at http://homepage.tudelft.nl/19j49), with default parameters. We also use CONNvis with default parameters automatically computed from the data statistics to produce coarse clusters.

First, we use the three simple datasets in Fig. 2, for which CONNvis clearly displays the cluster separations. Figs. 5, 6, 7 show the mappings for the Lsun, the Wingnut, and the Chainlink datasets, respectively. For the 2D datasets Lsun and Wingnut, Sammon's mapping and PCA provide a linear mapping with no effect on the visualization of separation boundaries, whereas for the entangled clusters of the 3D Chainlink they fail to express the cluster separation in the 2D projection. Similarly, while kernel PCA and SNE separate the clusters of the Lsun and Wingnut data, they are unsuccessful for the Chainlink data. Due to the fact that Isomap and LLE are focused on single-underlying manifold based on local neighborhood, they can map the well-separated clusters of the Lsun and the Chainlink as disconnected components, and thus they can successfully embed each component separately into 2D space. Even though separate mapping of disconnected components — clusters— may not show their similarity relations, it helps in cluster visualization. However, it is also possible to embed all clusters at once, by setting the neighborhood parameter (k = number of nearest data neighbors) accordingly (i.e., larger than the number of data samples in a cluster) to obtain one connected component of all data samples and have it projected as in Figs. 5, 7. For the inhomogeneously distributed Wingnut

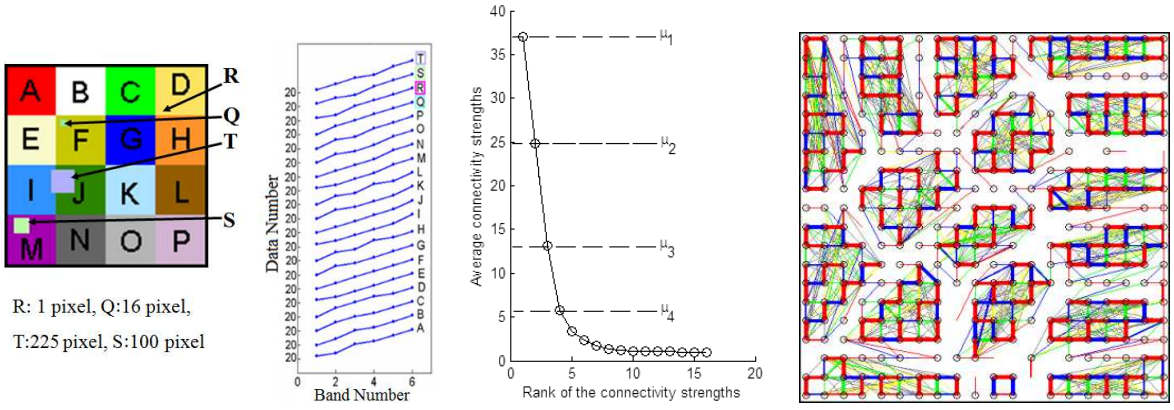


(a) Lsun (b) CONNvis in the data space (c) CONNvis on the SOM

(d) Wingnut (e) CONNvis in the data space (f) CONNvis on the SOM

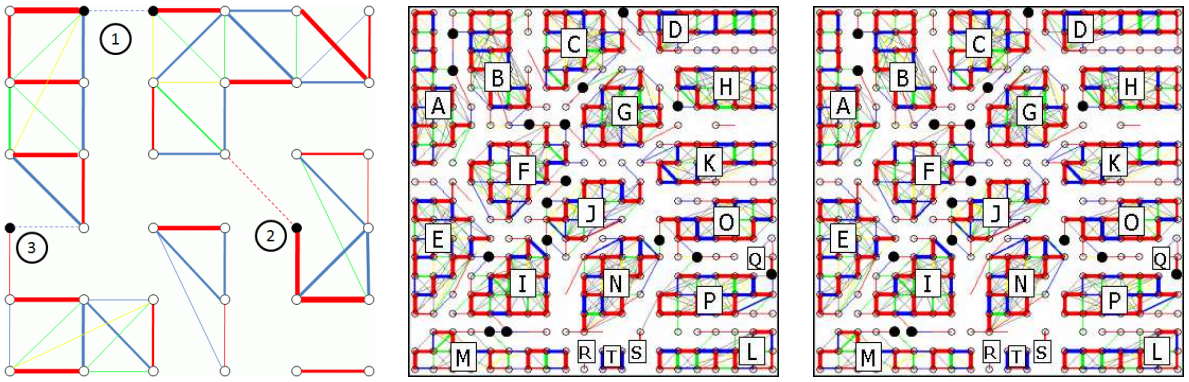
(g) 3D Chainlink (h) CONNvis in the 3D data space (i) CONNvis on the SOM

Figure 2: Example CONN visualization (CONNvis) in the data space and on the SOM, for three simple datasets from (Ultsch, 2005). A 10×10 SOM is used to obtain prototypes. Data points are shown as small grey dots in the left and middle columns, prototypes are labeled with unique symbols according to their known clusters. Top: (a) Lsun (2D data with three clusters) and its prototypes with true labels in data space. (b) CONN of Lsun prototypes, visualized in the data space. (c) CONNvis of Lsun prototypes on the SOM. Middle: (d) Wingnut dataset (two 2D clusters with inhomogeneous density distribution) and SOM prototypes (e) CONN of Wingnut prototypes in the data space. (f) CONNvis of Wingnut prototypes on the SOM. Bottom: (g) 3D Chainlink data (two linked 2D rings). (h) CONNvis of Chainlink prototypes in the 3D space (i) CONNvis on the SOM grid. The clusters of Lsun, Wingnut and Chainlink can be seen using CONNvis, through lack of connections (empty corridors) between clusters. (Figure is in color on-line.)



(a) 20-class, 6D data (b) Connectivity statistics (c) CONNvis on the SOM

Figure 3: (a) Left: the spatial distribution of the data vectors in the 128×128 pixel 6D synthetic image. Four clusters are relatively small, one (R) has only one pixel. Right: the mean signatures of the 20 classes, vertically offset for clarity. (b) Mean connectivity strengths (indicated by $\mu_1 - \mu_4$) for each rank of the first four ranks of the connections. (c) CONNvis, obtained by the 4-level binning scheme based on the statistics of the connections. Coarse clusters (strongly connected groups of prototypes) can be detected despite many (weak) topology violations, shown as connections between prototypes that are not neighbors in the SOM lattice. Because of the connectivity strengths, one gets an immediate sense of the relative importance of topology violations caused by a large or small number of data points. For cluster capture, topology violations that occur within clusters are inconsequential. More details on the data, CONN and CONNvis in the example are in (Merényi et al., 2007b) and in (Taşdemir and Merényi, 2009). (Figure is in color on-line.)



(a) (b) (c)

Figure 4: Illustration of interactive clustering from CONNvis. Some groups of prototypes — coarse clusters — are already outlined through the lack of connections when the visualization starts. The prototypes at the cluster boundaries, determined by visual inspection of the coarse clusters and the connections across them, are shown by black dots. (a) Circled numbers indicate three situations in interactive clustering described in Section 3.1. 1: different numbers of connections to each neighboring coarse cluster. (The prototype at the boundary of the cluster on the left has 4 within-cluster and 1 between-cluster connections, whereas the other —the one on the right— has 3 within-cluster and 1 between-cluster connections.) 2: the same number of connections to each cluster with different connectivity strengths 3: the same number of connections to each cluster with the same strengths but different rankings. The connections to be removed are drawn by dashed lines. (b) CONNvis of the 20-class data set (with weak global violations ($length > 2$, $strength < \mu_4$) removed). Letters indicate the truth labels of clusters. (c) Clusters resulting from the interactive clustering (by removing connections of the prototypes at the cluster boundaries). All clusters, including the one-pixel cluster R, are captured. (Figure is in color on-line.)

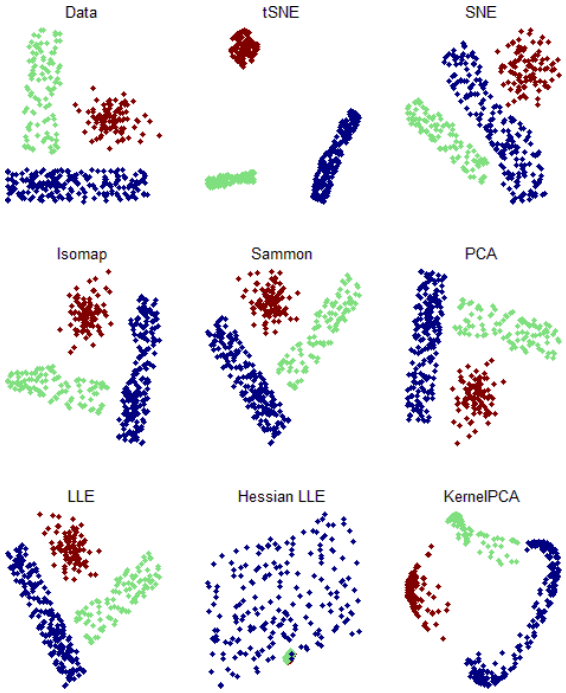


Figure 5: Visualization of the 2D Lsun dataset, shown with truth labels at the upper left, by eight different dimension reduction methods. Each project the data into 2D space. All but the Hessian LLE and Kernel PCA do a very good job separating the three clusters. tSNE stands out by exaggerating the separations. For 2D datasets, Sammon's mapping and PCA are linear transformations, with no effect on the visual separation of clusters. (Figure is in color on-line.)

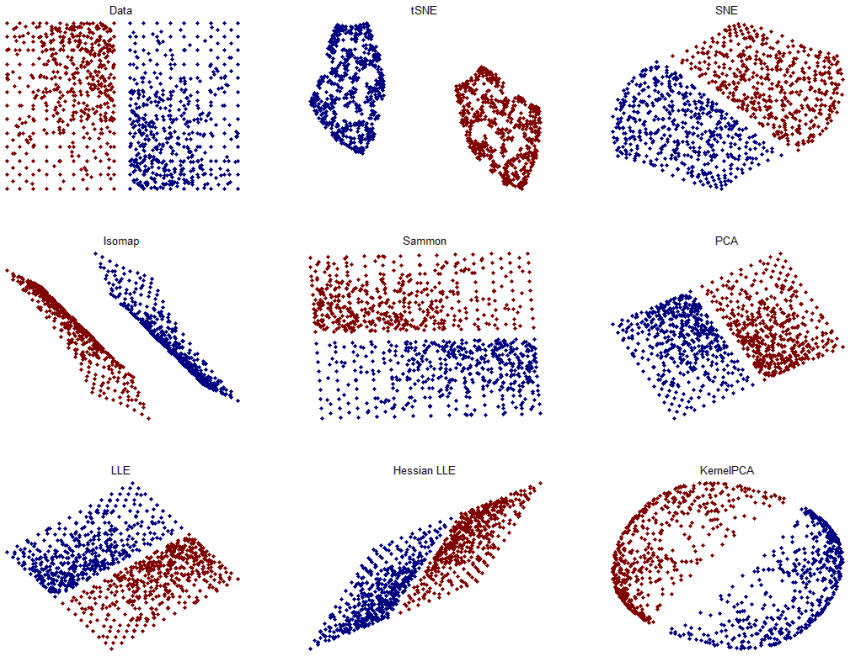


Figure 6: Visualization of the 2D Wingnut dataset (shown at upper left, with truth labels superimposed), in 2D space, by the same eight methods as in Fig. 5. While all methods provide perfect visual separation, the gaps tSNE and Isomap produce are larger than any within-cluster distances thus could also be used for automated extraction of the clusters. (Figure is in color on-line.)

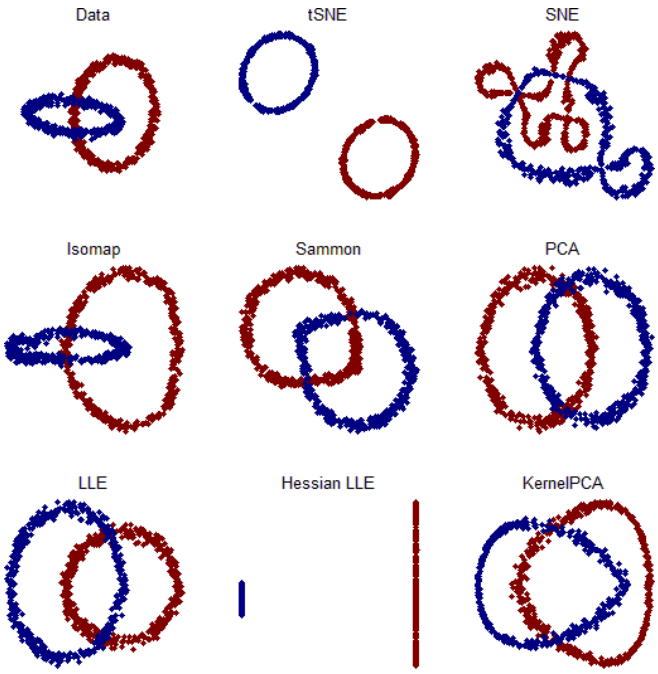


Figure 7: Visualization of the 3D Chainlink dataset (at upper left, with truth labels), through projection to 2D space by the same methods as in Figs. 5, 6. (Note that the data is visualized in 3D space). tSNE, LLE and the Hessian LLE map two clusters separately, whereas Sammon's mapping, PCA, Kernel PCA and SNE fail to produce visual delineation of these clusters with nonlinear separation boundary. See more explanation in the text. (Figure is in color on-line.)

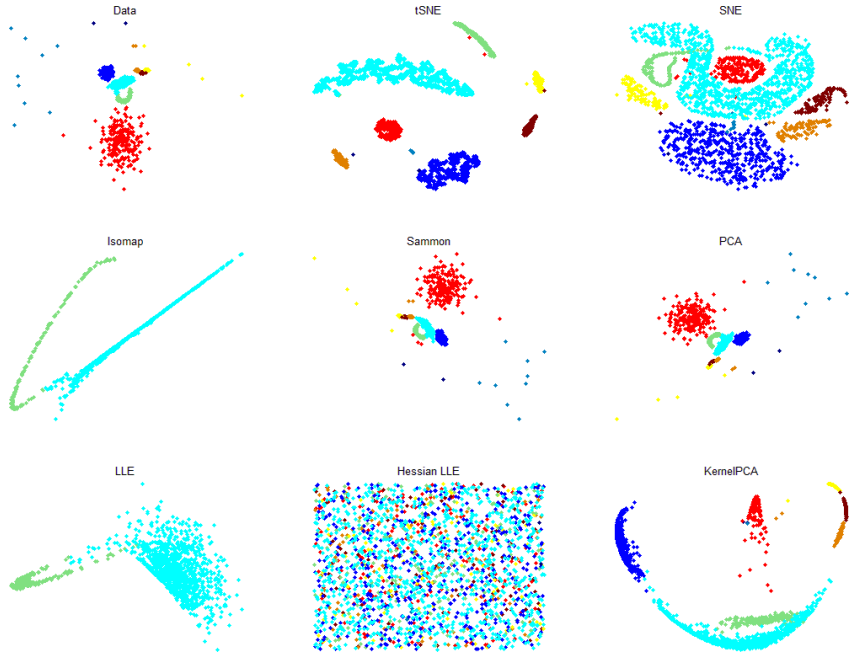


Figure 8: Visualization of the 2D Clown dataset by projection to 2D space. This dataset has more clusters (with more varying statistics) than the previous ones. tSNE visually separates the clusters with some points projected incorrectly apart from their true classes (red dots next to the green cluster) and with extra delineation of a cluster. (In some runs it partitions the cyan cluster into sub-parts, which do not exist in the data.). Neither the linear transformations of Sammon's mapping and PCA, nor the nonlinear projections of kernel PCA, SNE, LLE and Isomap produce a better visual delineation of the cluster structure. For Isomap and LLE, only the mapping of the largest connected component is displayed.

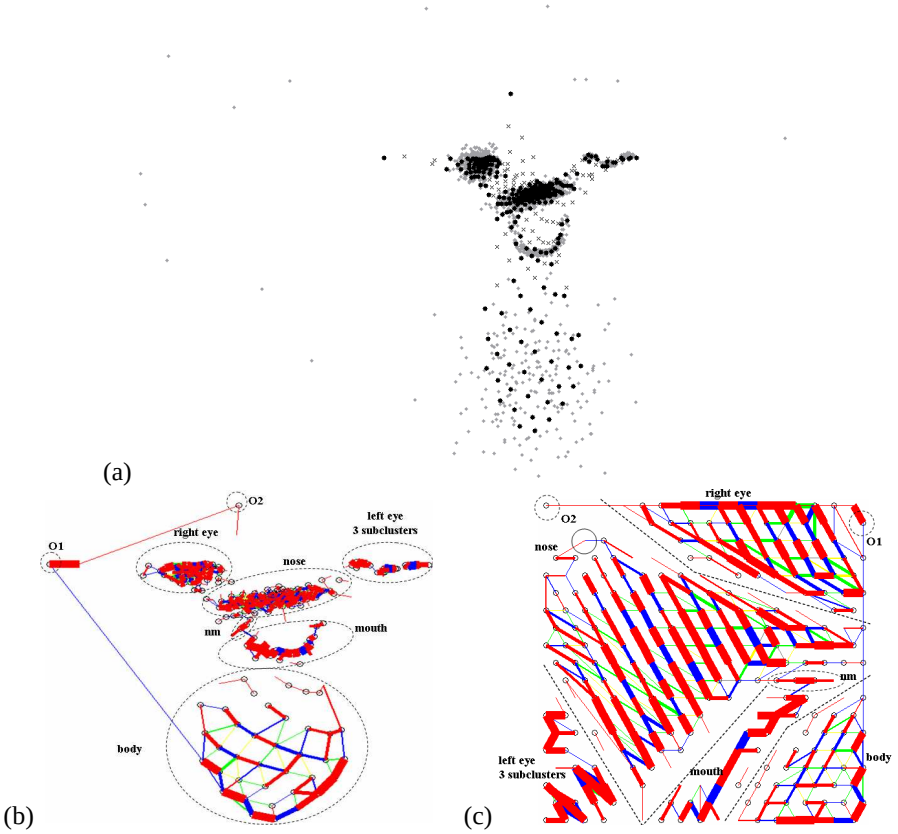


Figure 9: (a) The Clown dataset by Vesanto and Alhoniemi (2000) consisting of 2220 data points (gray dots) and its 19×17 SOM prototypes (graciously provided by Dr. Alhoniemi). (b) CONNvis in the data space. Prototypes are labeled with different symbols according to their clusters. (c) CONNvis on the SOM. (Figure is in color on-line.)

clusters whose distance is smaller than some of the within-cluster distances, (i.e., close enough to construct the neighborhood graph as one connected graph), the use of geodesic distances in Isomap helps better delineation of the boundary. For all these three datasets, tSNE not only displays the clusters separately but also increases the visual separation between them, due to its locally optimal cost function.

Second, we use another 2D dataset, the Clown (Fig. 8, upper left) constructed by Vesanto and Alhoniemi (2000). The Clown dataset has several clusters of varying statistics (different shapes, sizes and densities): an elliptical nose (turquoise), a u-shaped mouth (green), a spherical eye (blue), three small clusters (brown, orange, and yellow) representing the other eye, and a sparse body (red); and some outliers. Fig. 8 also shows the different mappings of the Clown dataset. Due to the mapping from 2D to 2D, Sammon's mapping and PCA are neither better nor worse than the original representation in terms of visual cluster separation, although Sammon's mapping changes the spatial relationship of the mouth and body. Isomap and LLE produce disconnected components, where the largest one consists of the two largest clusters (turquoise and green, the nose and the mouth) of the Clown dataset. Kernel PCA slightly improves the separation between disconnected parts of the dataset, producing a relatively more intelligible delineation than the original representation. Despite reordering of the clusters in a more compact layout (due to crowding problem discussed in van der Maaten and Hinton (2008)), SNE keeps the visual separation between the clusters and improves delineation of three small ones (orange, yellow and brown). By addressing the crowding problem in SNE, tSNE maps all clusters apart, at the expense of extra separations for relatively large clusters in some runs. Although our focus in this paper is on the evaluation of cluster separation in visualizations, we note that, despite the fact that the mapping is from 2D to 2D, the topological ordering of the clusters is preserved only by the PCA and the Kernel PCA. The CONNvis representation of similarities between quantization prototypes, shown in Fig. 9 clearly indicates the different clusters in the Clown dataset, as well as preserves the topology. (Eyes are on either sides of the nose, the mouth between the nose and body, like in the data set.)

Third, we map the synthetic 6D 20-class data into 2D space (Fig. 10). Due to the computational complexities of the dimensionality reduction methods, it is infeasible to use all 16384 data points. Therefore, using a priori class information, a quarter of the data points is randomly selected from each class (except for the 1-pixel class) to reduce the number of data points while preserving the respective ratio of the class sizes. Most methods (PCA, LLE, Isomap, and SNE) are successful in mapping these classes onto separate points. This is due to the Gaussian distribution of the data points in the classes of this synthetic data set. Sammon's mapping, however, obtains additional subcluster structures to reduce the mapping error in large clusters; whereas kernel PCA lumps data points into twelve mapping points (loses eight clusters). Surprisingly, tSNE is unsuccessful in projecting the 20 classes separately, despite the fact that the classes can be visually delineated by traditional approaches.

Finally, we study the projections of a real dataset, an 8D remote sensing image of Ocean City (Csathó et al., 1998). This dataset has 262,144 data points (512×512 pixel image) where each point is an 8D vector (a measured surface radiance in eight spectral band passes). In earlier work 28 clusters of widely varying statistics (including several relatively rare clusters) were identified and verified in this dataset (Merényi et al., 2007a, 2009). We gauge the performances of the dimensionality reduction methods and CONNvis against these verified earlier results. Due to the computational challenges, we select 200 data points per class with stratified random sampling using a priori class information. This makes the comparisons a little less than fair for CONNvis since the other methods are given equal-sized clusters, and also considerably fewer data points to map.

Fig. 11 shows the projections with superimposed class labels known from previous studies. While the majority of the dimensionality reduction methods group the data points reasonably well in the respective projections, from most of them it is not possible to discriminate separation boundaries among classes without the a priori class labels. Thanks to its cost function, emphasizing local similarities in a way that helps address the crowding problem, tSNE is considerably more successful than the other methods, for visual cluster identification in this dataset. As seen in Fig.11, it produces disconnected groups that indicate major dissimilarities among the data points. However, even this representation can distinguish only some groups. (We note that, due to its cost function aiming to reach a local minimum, tSNE may produce different mappings at each run. Therefore, orientation of clusters and splitting of subclusters may differ at each mapping; however, the same result appears at all runs: tSNE indicates major dissimilarities at the expense of unnecessary splitting of some clusters.)

Fig. 12 shows in more detail that each of these tSNE groups comprises several of the known clusters which are more similar to one another than to clusters contained in other groups. As an example, the boomerang-shaped contiguous set of points in the center represents a smooth change-over across five clusters, starting with the dark green cluster L and continuing on the diagonal toward the upper right with clusters O (split-pea green), N (orange), P (brown), and Q (ocher). While no boundaries can be discerned, the spectral signatures in Fig. 15 confirm that these clusters are not only similar but their similarity relations are correctly reflected by the tSNE layout. A similar case can be made for other (smaller) groups such as (A, j, G) or (I, R, J). It is also easy to see that the spectral signatures of the clusters within the (L,O,N,P,Q) group are more similar to one another than to those in clusters in the (A, j, G) and (I, R, J) groups. We leave it to the reader to inspect that the same holds for other groups.

An opposing tendency is also revealed in the tSNE map. Some known clusters are split and assigned into two or more groups at different locations. For example, cluster D (hot pink) appears near the top center and also as part of the (C, V, X, D) group at the bottom; cluster E (light blue) is an appendix of the (A, j, G) group and also appears near other clusters toward

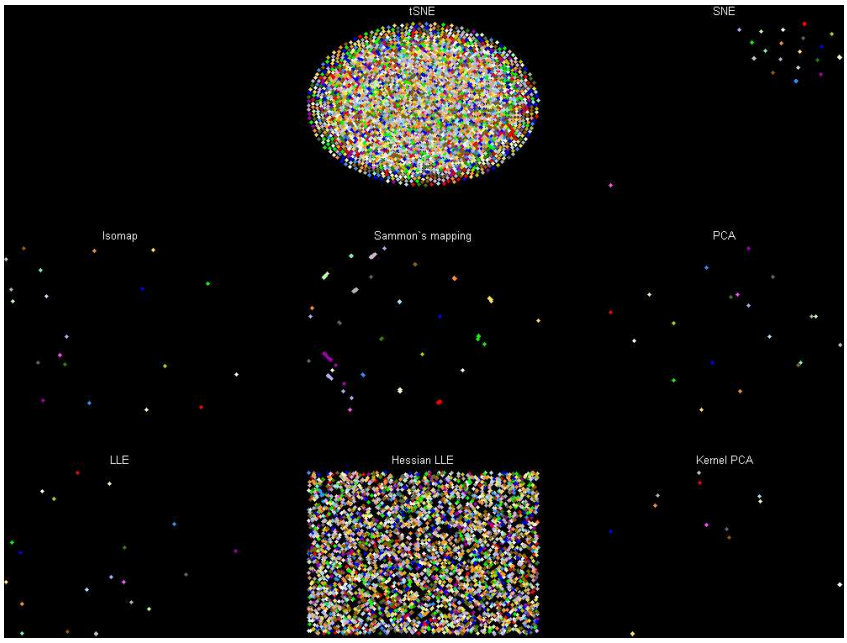


Figure 10: Visualization of the 6D 20-class dataset in 2D by projection. 25 % of the data points are selected from each class except for the 1-pixel class) to keep MATLAB processing time manageable. Due to the Gaussian construction of this synthetic dataset, PCA, Kernel PCA, and SNE map each class to a single point in the 2D projection. In addition, SNE provides a clear separation of the 1-pixel class (with significantly different mean feature vector from others as shown in Fig. 3), by mapping it far from the others. Surprisingly, tSNE does not produce a clear separation. (We note, however, that when 1-pixel class is removed, tSNE maps remaining classes to 19 points representing the corresponding classes.)

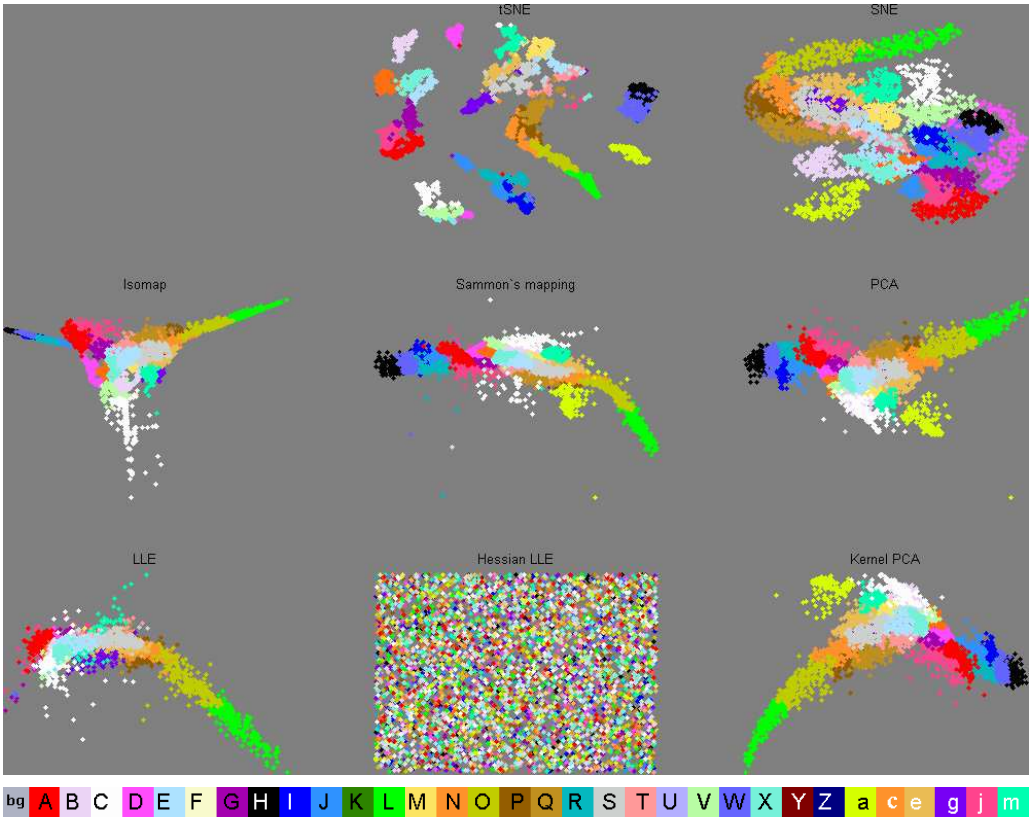


Figure 11: Visualization of the 8D Ocean City data in 2D projection. 200 randomly selected points are used from each of the 28 classes, because of computational constraints on the MATLAB toolbox. The truth labels are shown as different colors. tSNE provides some separation among clusters, however, none of these dimensionality reduction methods provides detailed visualization for interactive capture of all known groups in the data. (Figure is in color on-line.)

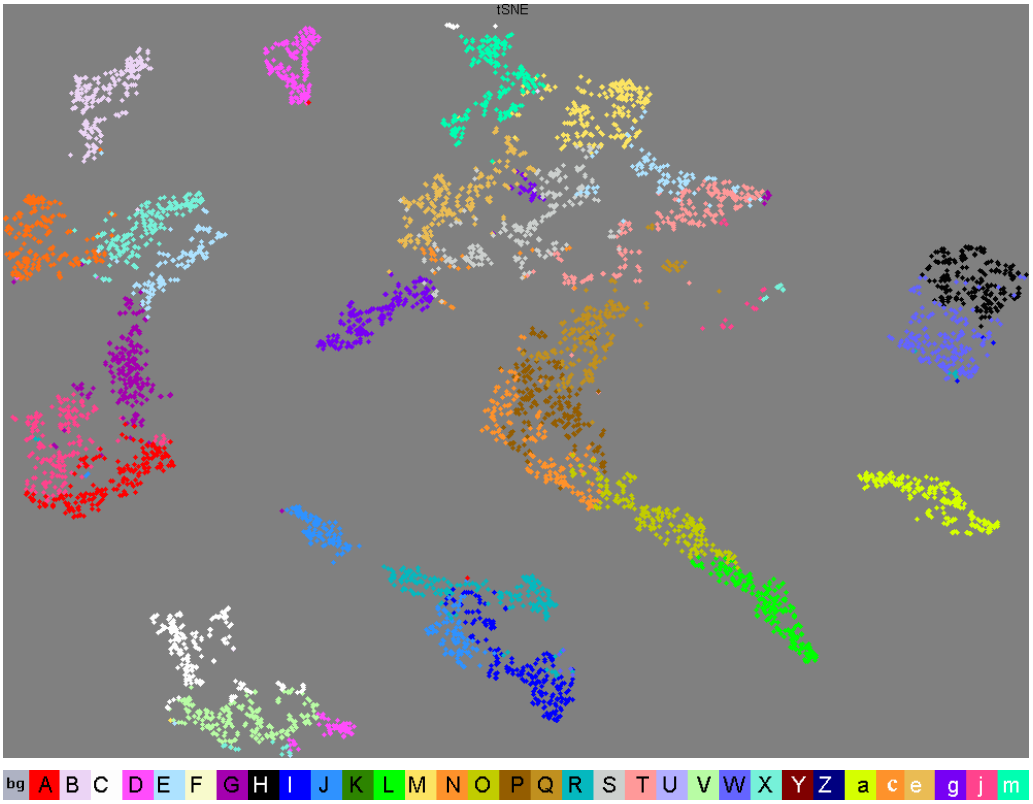


Figure 12: A magnified version of the tSNE projection of the 8D Ocean City data in 2D, from Fig. 11. (Figure is in color on-line.)

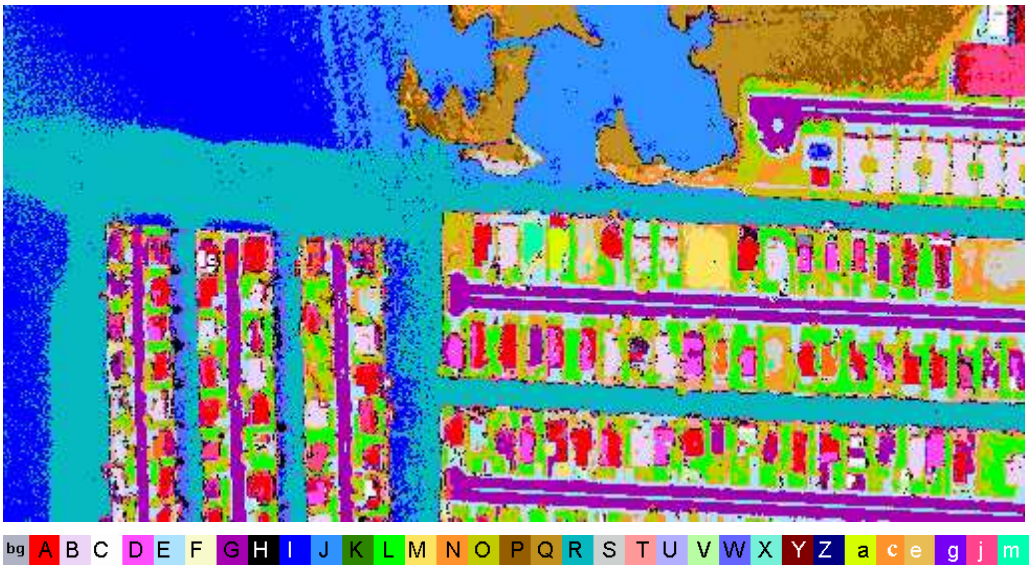


Figure 13: 28 clusters of the 8D Ocean City dataset identified from SOM clustering and shown over the spatial image. The complete image with the known clusters can be seen in (Merényi et al., 2007b) and in (Merényi et al., 2007a). (Figure is in color on-line.)

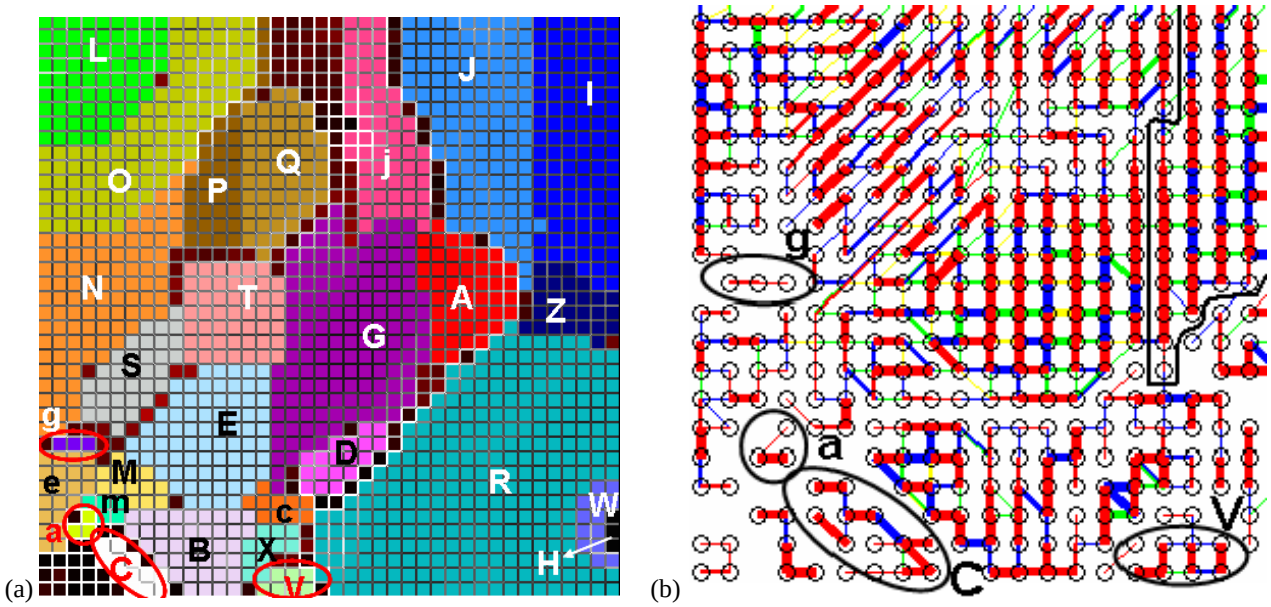


Figure 14: (a) Clusters of the Ocean City image identified from CONNvis visualization of the SOM (colored groups of cells), and cluster labels (letters) shown. (b) The CONNvis of the lower left quadrant (for reasons of space limitations) of the SOM at left. Several obvious clusters, indicated by detached islands of protoypes, are circled or delineated by lines, for examples of cluster extraction. (Figure is in color on-line.)

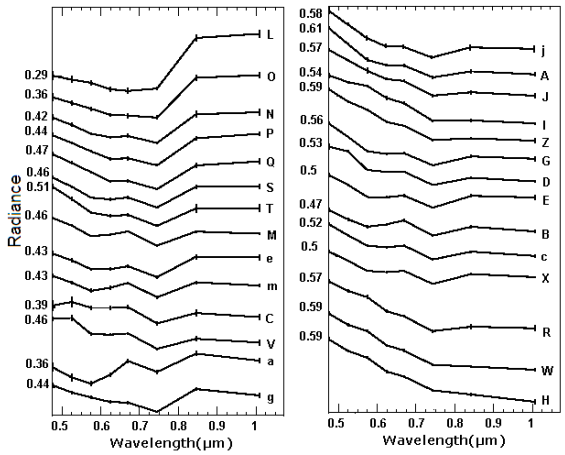


Figure 15: Mean spectral signatures of the 28 clusters extracted from the 8D Ocean City dataset. Clusters are excellent match to those verified in earlier work (Merényi et al., 2007a,b) and therefore we show the mean spectra of CONNvis clusters here. Spectra are vertically offset for clarity. The numbers on the y axis indicate the radiance value of each spectrum in the first spectral band. The small vertical tick marks on the spectra show the standard deviation of the respective classes in the corresponding band passes. (Figure is in color on-line.)

the top center (between M (yellow) and T (salmon)). Clusters G (magenta), C (white), g (purple) are further instances of split clusters. As the standard deviations of the clusters is quite small (as indicated in Fig. 15, especially for clusters (E, D, G, g), substantially different subclusters, warranting the splits are unlikely. Perhaps more interestingly, we can see two groups, (H, W) and (J, R, I), which contain clusters with similar signatures but are mapped at very different locations.

In summary, tSNE produces a sensible grouping of the data. However, even this representation can separate only eight groups, and it seems to overemphasize some possibly very slight differences, which may result in undue splitting of clusters.

From a CONNvis representation we were able to identify the 28 known clusters. We show, in Fig. 13, part of the spatial image with the pixels color-coded according to the identified surface material clusters. The corresponding SOM and a magnified portion of the CONNvis visualization of the corresponding SOM is in Fig.14. On the SOM the colors show the different identified clusters, while the letters give the corresponding labels to tie with the mean cluster spectra in Fig. 15. Because of space constraints we only show the lower left quadrant of the CONNvis. Discontinuities in the data space are expressed by the lack of connectivity in the CONNvis between adjacent SOM grid locations. In Fig.14(b), completely isolated groups of prototypes are obvious (several are circled or delineated by lines). These helped immediate identification of clusters such as V (at the lower right of the CONNvis detail), C, a, or g. The full images and more details about the CONNvis mapping including comparison with earlier results can be found in (Taşdemir and Merényi, 2009; Merényi et al., 2007a).

4.1. Computational complexity

We note that all dimensionality reduction methods (used in this study) are seriously limited in their ability to handle large datasets, due to their computational complexities and memory requirements. As an evidence of this, we had to subsample even relatively small datasets such as the 6D synthetic image cube, which contains 16,386 data vectors. Our real 8D remote sensing spectral image had to be subsampled even more severely: 28 (classes) $\times$ 200(= 5600) points were used out of

$512 \times 512 = 256,000$ points. In general, Isomap, Sammon's mapping, and kernel PCA have a computational complexity of $O(n^3)$ (n is the number of data samples), with a memory requirement of $O(n^2)$; whereas LLE and Hessian LLE have a computational complexity of $O(r_p n^2)$ ($r_p \leq 1$ is the ratio of positive elements in a sparse matrix to the total number of elements), with a memory requirement of $O(r_p n^2)$ (van der Maaten et al., 2009). SNE and tSNE have similar limitation, due to their computational complexity and memory requirement of $O(n^2)$ (van der Maaten and Hinton, 2008). CONNvis, in contrast, can handle large data volumes easily, due to its relatively low computational load — $O(i * n * n_w)$ (i is number of iterations, n_w is the number of SOM neural units, which is considerably smaller than n for large datasets)— and significantly less memory requirement of $O({n_w}^2)$.

5. Discussion and conclusions

CONNvis is a 2D graph-based visualization of datasets, based on *i)* the spatially ordered set of quantization prototypes obtained by an SOM, and *ii)* prototype similarities expressed by a topology representing CONN graph (a weighted version of induced Delaunay graph). CONNvis thus enables informative visualization of prototype similarities defined by detailed local data distribution. Experiments on various datasets (with different dimensionalities and with varying cluster statistics) indicate that CONNvis is a successful 2D visualization for interactive interpretation of complex data structures.

Among dimensionality reduction methods used in this study, only tSNE comes close to CONNvis for expressive visualization of high-dimensional data with complex structure. The relative success of tSNE is mainly due to its cost function emphasizing local similarities in the projected space using a heavy-tailed distribution to compensate the dimensionality mismatch (i.e., high-dimensional data space versus its low-dimensional projection). Due to the same reason, tSNE may exaggerate slight differences within clusters, which in turn may produce unnecessary splitting of some natural clusters.

Except for CONNvis, all methods discussed in this paper visualize data directly in the data space or in a 2D projection of the data space. More precisely, they visualize the distances of data points or projections of data points according to a metric distance defined in data space. This limits optimal viewing to data that are no more than 3D, or data of higher dimensions containing linearly separable spherically symmetrical clusters for which projections do not hide structural details. CONNvis is fundamentally different because the visualization for higher than 2-3D data is done in a representation space — the SOM lattice — translating the data space distances to separation or connection strengths, draped over the topologically ordered lattice of the SOM protoypes. While SOM visualization loses the sense of the customary (Euclidean) distances between prototypes it is more expressive of the structure (submanifolds, clusters) of complicated high-D spaces than visualizations in data space. One might argue that due to the prototypical representation SOM visualization also loses some detail existing in the data. While this is true, SOM learning — specifically the Conscience SOM that we use — transfers the maximum information content from data to (the given number of) prototypes. With precise evaluation and monitoring the quality of SOM learning details important for the characterization of the manifold structure can be separated from the unimportant (s.a. noise), thus critical information preserved and properly summarized (Merényi et al., 2009) to a manageable size. CONNvis further enriches previous SOM visualizations by summarizing data characteristics on a sub-prototype level (instead of prototype level) and bringing the local, unisotropic distribution information to the "surface, for more nuanced delineation of clusters than SOM visualizations showing only the dissimilarities of the SOM-neighbor prototypes and / or the size of their receptive fields (as in U-matrix type visualizations). It also shows the topological relations of all prototypes, not only for neighbors in the SOM lattice, unlike U-matrix type representations. These properties considerably enhance the representation of data structure, which is the primary goal of exploratory visualization. We recognize that CONNvis is more difficult to look at and to interpret than (for example) the U-matrix, and many other visualizations. However, the U-matrix and variants, which are routinely used today, also took some time for the community to get used to. Data mining needs increasingly expressive tools for the navigation of high-D, complex data sets. In this era of "big data" visualizations that can also scale with the data volume fulfill an additional critical demand. For these reasons we believe that effort invested in understanding the CONNvis representation and visualization controls has a rewarding pay-off.

Acknowledgements

Many thanks to Prof. B. Csathó, from the Department of Geology, University of Buffalo, for the Ocean City image and ground truth. We also thank Dr. Alhoniemi, Department of Information Technology, Turku University, for the Clown data and the 19×17 SOM prototypes. Some of the SOM data analyses and development of software tools by the Neural Machine Learning Group at Rice University were partially supported by NASA grants NNG06GE95G, NNG05GA94G.

References

Aupetit, M., 2006. Learning topology with the Generative Gaussian Graph and the EM algorithm, in: Weiss, Y., Schölkopf, B., Platt, J. (Eds.), Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA, pp. 83–90.

Bishop, C.M., Svensen, M., Williams, C.K.I., 1998. GTM: The generative topographic mapping. Neural Computation 10, 215–234.

Cottrell, M., B, H., Hasenfuss, A., Villmann, T., 2006. Batch and median neural gas. Neural Networks 19, 762–771.

Cottrell, M., de Bodt, E., 1996. A Kohonen map representation to avoid misleading interpretations, in: Proc. 4th European Symposium on Artificial Neural Networks (ESANN'96), Bruges, Belgium, D-Facto, pp. 103–110.

Cox, T.F., Cox, M., 2001. Multidimensional Scaling. Chapman and Hall/CRC.

Csathó, B., Krabill, W., Lucas, J., Schenk, T., 1998. A multisensor data set of an urban and coastal scene, in: Intl Archives of Photogrammetry and Remote Sensing, pp. 26–31.

DeSieno, D., 1988. Adding a conscience to competitive learning, in: IEEE International conference on Neural Networks, pp. I–117–I–124.

Donoho, D.L., Grimes, C., 2003. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, in: Proc. of the National Academy of Sciences, pp. 5591–5596.

Hakkinen, E., Koikkalainen, P., 1997. The neural data analysis environment, in: Proc. 1st Workshop on Self-Organizing Maps (WSOM'97), Espoo, Finland, June 4-6, pp. 69–74.

Hamel, L., Brown, C., 2011. Improved interpretability of the unified distance matrix with connected components, in: Proc. 7th International Conference on Data Mining (DMIN'11), CSREA Press, Las Vegas. pp. 338–343.

Hammer, B., Hasenfuss, A., Villmann, T., 2007. Magnification control for batch neural gas. Neurocomputing 70, 1125–1234.

Himberg, J., 2000. A SOM based cluster visualization and its application for false colouring, in: Proc. IEEE-INNS-ENNS International Joint Conf. on Neural Networks, Como, Italy, pp. 587–592.

Hinton, G.E., Roweis, S.T., 2002. Stochastic neighbor embedding, in: Advances in Neural Information Processing Systems 15, MIT Press. pp. 833–840.

Kaski, S., Kohonen, T., Venna, J., 1998. Tips for SOM processing and colour-coding of maps, in: G. Deboeck, T.K. (Ed.), Visual Explorations in Finance Using Self-Organizing Maps. Springer, London.

Kaski, S., Venna, J., Kohonen, T., 2000. Coloring that reveals cluster structures in multivariate data, in: Australian Journal of Intelligent Information Processing Systems, pp. 82–88.

Kohonen, T., 1997. Self-Organizing Maps. Springer-Verlag Berlin Heidelberg. 2nd edition.

Kraaijveld, M., Mao, J., Jain, A., 1995. A nonlinear projection method based on Kohonen's topology preserving maps. IEEE Trans. on Neural Networks 6, 548–559.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research , 2579–2605.

van der Maaten, L., Postma, E., van den Herik, H., 2009. Dimensionality Reduction: A Comparative Review. Technical Report TiCC-TR 2009-005. Tilburg University Technical Report.

Martinetz, T., Berkovich, S., Schulten, K., 1993. Neural gas network for vector quantization and its application to time series prediction. IEEE Trans. Neural Networks 4, 558–569.

Martinetz, T., Schulten, K., 1994. Topology representing networks. Neural Networks 7, 507–522.

Merényi, E., 2000. "Precision Mining" of high-dimensional patterns with Self-Organizing Maps: Interpretation of hyperspectral images, in: Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence. Studies in Fuzziness and Soft Computing, Physica-Verlag.

Merényi, E., Csathó, B., Taşdemir, K., 2007a. Knowledge discovery in urban environments from fused multi-dimensional imagery, in: Proc. 4th IEEE GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (URBAN 2007), P. Gamba and M. Crawford, Eds., Paris, France, 11-13 April 2007, IEEE Catalog number 07EX1577.

Merényi, E., Jain, A., Villmann, T., 2007b. Explicit magnification control of self-organizing maps for "forbidden data". IEEE Trans. on Neural Networks 18, 786–797.

Merényi, E., Taşdemir, K., Zhang, L., 2009. Learning highly structured manifolds: harnessing the power of soms, in: Biehl, M., Hammer, B., Verleysen, M., Villmann, T. (Eds.), Similarity based clustering, Lecture Notes in Artificial Intelligence. Springer-Verlag,. volume 5400, pp. 138–168.

Merkl, D., Rauber, A., 1997. Alternative ways for cluster visualization in Self-Organizing Maps, in: Proc. 1st Workshop on Self-Organizing Maps (WSOM'05), Espoo, Finland, June 4-6. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, pp. 106–111.

Pampalk, E., Rauber, A., Merkl, D., 2002. Using smoothed data histograms for cluster visualization in Self-Organizing Maps, in: Proc. of Int'l Conference on Artificial Neural Networks-ICANN 2002,Madrid, Spain, August 28-30, pp. 871–876.

Polito, M., Perona, P., 2001. Grouping and dimensionality reduction by locally linear embedding, in: Advances in Neural Information Processing Systems 14, MIT Press. pp. 1255–1262.

Ritter, H., Schulten, K., 1986. On the stationary state of Kohonen's self-organizing sensory mapping. Biol. Cyb. 54, 99–106.

Roweis, S., Soul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Su, M.C., Chang, H.T., 2001. A new model of self-organizing neural networks and its applications. IEEE Transactions on Neural Networks 12, 153–158.

Taşdemir, K., Merényi, E., 2009. Exploiting data topology in visualization and clustering of Self-Organizing Maps. IEEE Transactions on Neural Networks 20, 549–562.

Tenenbaum, J.B., de Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Ultsch, A., 1993. Self-organizing neural networks for visualization and classification, in: Lausen, O.B., Klar, R. (Eds.), Information and Classification-Concepts, Methods and Applications. Springer Verlag, Berlin, pp. 307–313.

Ultsch, A., 2003. Maps for the visualization of high-dimensional data spaces, in: Proc. 4th Workshop on Self-Organizing Maps (WSOM'03), pp. 225–230.

Ultsch, A., 2005. Clustering with SOM: U*c, in: Proc. 5th Workshop on Self-Organizing Maps (WSOM'05), Paris, France, September 5-8, pp. 75–82.

Vesanto, J., 1999. SOM-based data visualization methods. Intelligent Data Analysis 3, 111–126.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11, 586–600.

Villmann, T., Claussen, J., 2005. Magnification control in self-organizing maps and in neural gas. Neural Computation 18, 446–469.

Villmann, T., Merényi, E., 2001. Extensions and modifications of the Kohonen SOM and applications in remote sensing image analysis, in: Seiffert, U., Jain, L.C. (Eds.), Self-Organizing Maps: Recent Advances and Applications. Springer-Verlag, pp. 121–145.

Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G., Koudas, N., 2002. Non-linear dimensionality reduction techniques for classification and visualization, in: Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 645–651.

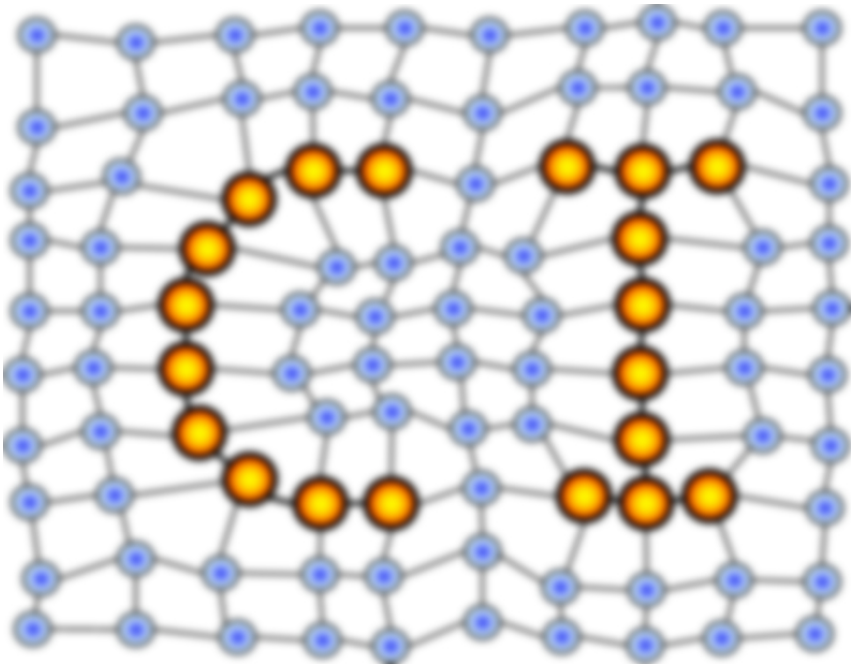
Yang, M.H., 2002. Face recognition using extended Isomap, in: Proc. of International Conference on Image Processing (ICIP2002), September 22-25, Rochester, NY, pp. 117–120.

Yin, H., 2002. ViSOM- a novel method for multivariate data projection and structure visualization. IEEE Transactions on Neural Networks 13, 237–243.

Zhang, J., Li, S., Wang, J., 2004. Manifold learning and applications in recognition.

MACHINE LEARNING REPORTS

Report 05/2012



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.