Learning Highly Structured Manifolds: Harnessing the Power of SOMs

Erzsébet Merényi, Kadim Tasdemir, and Lili Zhang

Department of Electrical and Computer Engineering, Rice University, Houston, Texas, U.S.A.

Abstract. In this paper we elaborate on the challenges of learning manifolds that have many relevant clusters, and where the clusters can have widely varying statistics. We call such data manifolds *highly structured*. We describe approaches to structure identification through self-organized learning, in the context of such data. We present some of our recently developed methods to show that self-organizing neural maps contain a great deal of information that can be unleashed and put to use to achieve detailed and accurate learning of highly structured manifolds, and we also offer some comparisons with existing clustering methods on real data.

1 The Challenges of Learning Highly Structured Manifolds

Data collected today are often high-dimensional due to the vast number of attributes that are of interest for a given problem, and which advanced instrumentation and computerized systems are capable of acquiring and managing. Owing to the large number of attributes that are designed to provide sophisticated characterization of the problem, the data acquired are not only high-dimensional but also highly structured, i.e., the data have many clusters that are meaningful for the given application. Examples are hyperspectral imagery of planetary surfaces or biological tissues, DNA and protein microarrays, data bases for business operations and for security screening. These types of data created new demands for information extraction methods in regard to the detail that is expected to be identified. For example, hyperspectral imagery affords discrimination among many materials such as individual plant species, soil constituents, the paints of specific makes of cars, or a large variety of roof and building materials, creating a demand to extract as many as a hundred different clusters from a single remote sensing image of an urban scene. These clusters can be extremely variable in size, shape, density and other properties as we illustrate below. Another demand arising from such sophisticated data is to differentiate among clusters that have subtle differences, as the ability to do so can enable important discoveries. These examples highlight challenges for which many existing clustering and classification methods are not well prepared.

There has been much research on manifold learning motivated by the idea that the data samples, even if they are high-dimensional, can be represented by a low-dimensional submanifold. Representing the data in low-dimensional (2-d or 3-d)

M. Biehl et al.: (Eds.): Similarity-Based Clustering, LNAI 5400, pp. 138–168, 2009.

spaces can also help visualize the data structure to guide the user for capturing the clusters interactively. A classical technique for dimensionality reduction is principal component analysis (PCA) which works well when the data points lie on a linear submanifold. However, real data often lie on nonlinear spaces. To find the nonlinear subspaces (manifolds), many methods have been introduced (see [1,2] for recent reviews), among them a number of manifold learning algorithms: multidimensional scaling (MDS) [3], Isomap [4], locally linear embedding (LLE) [5], Hessian LLE (hLLE) [6] are some. These methods may successfully be applied to data sets that are characterized by only a few parameters (such as the angle of rotation in a number of similar video and image data sets) [7]. However, the same are often suboptimal for clustering applications as shown in various papers [8,9,10], since they are developed for reconstruction of one underlying submanifold rather than for identification of different groups in the data. In order to make manifold learning algorithms effective for classification, various works extend them with the help of the class statistics. [8] uses Fisher linear discriminant analysis (LDA) with Isomap for face recognition. Similarly, [11] uses LDA with manifold learning algorithms for face and character recognition. [10] modifies Isomap and LLE so that both local and global distances are considered for better visualization and classification. However, the performance of the modified Isomap or the modified LLE is not very promising due to the same reconstruction objective (the reconstruction of one underlying manifold) as for Isomap and LLE. In clustering applications, the aim is to learn the cluster structure — where the clusters may lie in different submanifolds — rather than to find one underlying submanifold for the data. Therefore representation of the separation between clusters is of great interest but not so much the precise topography of the underlying manifold. This makes adaptive vector quantization algorithms – which show the data topography on the prototype level and aim to faithfully represent the local similarities of the quantization prototypes – well suited for clustering [12,13,14,15,16].

Adaptive vector quantization algorithms are either inspired by nature as in the case of self-organizing maps (SOMs) [12], derived as stochastic gradient descent from a cost function as in Neural Gas [13] and its batch version [14], or derived through expectation-maximization [15,16]. Variants of SOM, Neural Gas and batch Neural Gas, which use a magnification factor in quantization, were also introduced and analyzed to control the areal representation of clusters (e.g., the enhancement of small clusters), in the learning process [17,18,19,20].

Among adaptive vector quantization methods we focus on SOMs. Our motivation is that not only can SOMs demonstratedly find optimal placement of prototypes in a high-dimensional manifold (and through that convey knowledge of the manifold structure) but the ordered prototypes also allow interesting and in-depth knowledge representation, regardless of the input dimensionality, which in turn helps resolve a large variety of clusters in great detail. After a brief background on SOM learning in Sect. 2, we discuss aspects of SOM learning as related to large high-dimensional manifolds with many clusters: quantification of the quality of learning (topology preservation) in Sect. 3.1; representation of

the SOM knowledge, and extraction of clusters under these circumstances in Sect. 3.2, and we describe methods we developed in recent years. In Sect. 4 we present real data analyses, and offer conclusions in Sect. 5.

2 Learning Manifolds with Self-Organizing Maps

Self-Organizing Maps (SOMs) occupy a special place in manifold learning. They perform two acts simultaneously, during an iterative learning process. One is an adaptive vector quantization, which — assuming correct learning — spreads the quantization prototypes throughout the manifold such that they best represent the data distribution, within the constraints of the given SOM variant. (For example, the Conscience SOM produces an optimum placement of the prototypes in an information theoretical sense [21,18].) The other act is the organization (indexing) of the prototypes on a rigid low-dimensional grid, according to the similarities among the prototypes as measured by the metric of the data space. This duality makes SOMs unique among vector quantizers, and unique among manifold learning methods, because the density distribution — and therefore the structure — of a high-dimensional manifold can be mapped (and visualized) on a 1-, 2- or 3-dimensional grid without reducing the dimensionality of the data vectors. This, in principle, allows capture of clusters in high-dimensional space, which in turn facilitates identification of potentially complicated cluster structure that is often an attribute of high-dimensional data.

However, "the devil is in the details". The aim of this paper is to illuminate and quantify some of the details that are different for simple data and for complicated (highly structured) data, and to describe our contributions that alleviate certain limitations in existing SOM approaches (including the interpretation of the learned map) for highly structured manifolds.

For a comprehensive review of the SOM algorithm, see [12]. To briefly summarize, it is an unsupervised neural learning paradigm that maps a data manifold $M \subset \mathbb{R}^d$ to prototype (weight) vectors attached to neural units and indexed in a lower dimensional fixed lattice A of N neural units. The weight vector w_i of each neural unit i is adapted iteratively as originally defined by Kohonen [12]: Find the best matching unit i for a given data vector $v \in M$, such that

$$\|v - w_i\| \leq \|v - w_j\| \; \forall j \in A \tag{1}$$

and update the weight vector 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 a neighborhood function, often defined by a Gaussian kernel around the best matching unit w_i. Through repeated application of the above steps, the weight vectors become the vector quantization prototypes of the input space M.

This is an enigmatic paradigm: it has been studied extensively (e.g., [12,22,23,24,25]), yet theoretical results are lacking for proof of convergence and ordering for the general case. In principle, the SOM is a topology

preserving mapping, *i.e.*, the prototypes which are neighbors in A should also be neighbors (centroids of neighboring Voronoi polyhedra as defined in [26]) in M, and vice versa. When centroids of neighboring Voronoi polyhedra in M are not neighbored in the SOM lattice we speak of *forward topology violation* with a folding length k where k is the lattice distance between the prototypes in question. When two SOM neighbors are not centroids of adjacent Voronoi polyhedra in M a *backward topology violation* occurs [27]. Both conditions can be (and are usually) present at various stages of SOM learning, to various extent depending on the characteristics of the data and the SOM lattice. This is so even if the learning "goes well" (no twists develop in the map), and topology violations may not vanish with any amount of learning. One is therefore motivated to construct empirical and heuristic measures to quantify the quality of learning: the *degree* of faithfulness and maturity of the mapping, as it is meaningful for a given application. Measures of topology preservation, such as the Topographic Product, Topographic Error, and Topographic Function, have been proposed [28,29,27], which define perfect mapping in exact numerical terms. The Topographic Function [27] also shows topology violations as a function of the folding length, which gives a sense of how global or local the violations are on average. What these existing measures do not provide, however, is a clear sense of what the numbers mean for less than perfect learning: how far the score of an already usefully organized state of the SOM could be from perfect (zero for the Topographic Product and the Topographic Error); whether a numerical value closer to the perfect score necessarily means better organization; or which of two Topographic Functions express better organization.

The quality of learning can be viewed in relation to the goal of the learning. Two levels are easily distinguished: a) the learning of the topography (the density distribution); and b) the learning of the cluster structure, of a manifold for which the acceptable degree of topographic faithfulness can be quite different. Learning cluster structure does not require a very precise learning of the topography. Certain level of local topology violations in the SOM is tolerable and will not hinder the accurate extraction of clusters. To illustrate the point, consider the SOM learned with a relatively simple data set, in Fig. 1. The data set contains 8 classes, in a 6-dimensional feature space. The classes are apportioned such that four of them comprise 1024 data vectors each, two of them have 2048, and two have 4096 data vectors. Gaussian noise, about 10% on average, was added to the data vetors to create variations within the classes. The clusters detected by the SOM are delineated by the white "fences" we call the mU-matrix, and by the empty (black) prototypes, as explained in the figure caption. (We will elaborate on this particular knowledge representation more in Sect. 3.2.) Comparison with the known class labels (colors) superimposed in Fig. 1, right, makes it obvious that the SOM already learned the cluster structure perfectly. At the same time, one can examine the topology violations and conclude that the topography is far from having been learned perfectly. To show this we connected with black lines those prototypes which are not neighbors in the SOM grid but have adjacent Voronoi cells in data space. These lines express the forward topology violations,

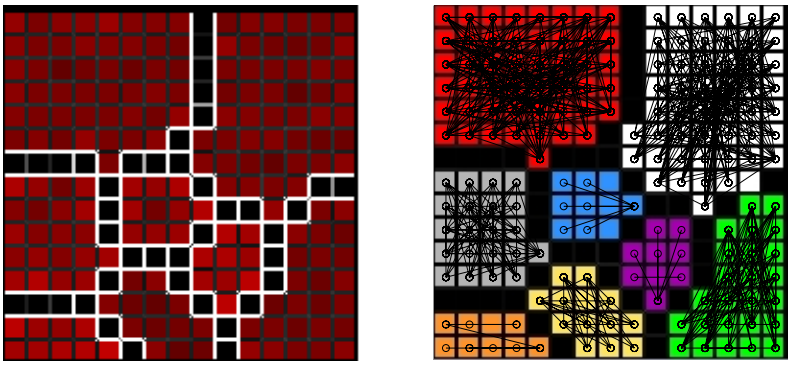

Fig. 1. Left: 15 x 15 SOM of an 8-class 6-dimensional synthetic data set. The cluster boundaries emerge through the mU-matrix, the visualization of the distances of adjacent prototypes as fences between the grid cells in all eight directions. White is high fence (large dissimilarity), black is low fence (great similarity). Each grid cell is shaded by an intensity of red proportional to the number of data points mapped to the prototype in that grid cell. Black grid cells indicate empty receptive fields of the corresponding prototypes. The fences clearly outline 8 clusters. **Right:** The SOM of the 8-class synthetic data with known class labels (colors) superimposed. The clustering learned by the SOM is perfect in spite that many topology violations still exist at this stage of the learning, even at folding length 8 (i.e., nearly half the width of the SOM lattice). The (567) existing violations, shown by black lines in the cluster map, all occur locally within clusters.

and we call them *violating connections*. All violating connections are residing locally inside clusters, without confusing the cluster boundaries. Apparently, a coarser organization has taken place, separating the clusters, and finer ordering of the protoypes would continue within the already established clusters. One might conclude that, for the purpose of cluster capture, this level of topology violation is tolerable and inconsequential [30]. We elaborate on aspects of measuring topology violations, and present new measures, in Sect. 3.1 and 3.2.

3 Learning the Clusters in Highly Structured Manifolds

As stated in Sect. 1 high-dimensional data are often complicated, highly structured, as a result of the application task for which the data were collected. Complicated means the presence of many clusters which may not be linearly separable, and which can be widely varying in various aspects of their statistics, such as size, shape (non-symmetric, irregular), density (some are very sparse, others are dense in feature space), and their proximities. Fig. 2 gives an illustration of these conditions.

To give a real example of a situation similar to that in Fig. 2 we show statistics of a remote sensing spectral image of Ocean City, Maryland [31]. This data set is described in detail in Sect. 4.1. In previous analyses more than twenty clusters

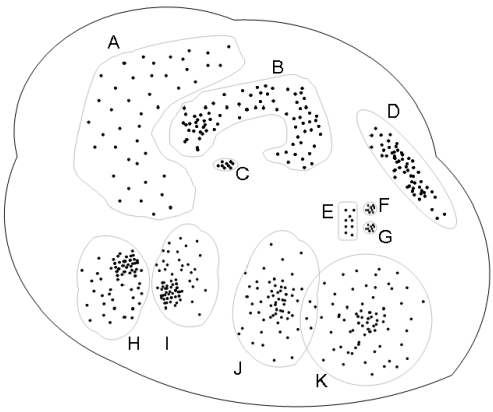

Fig. 2. Illustration of complicated cluster structure. The eleven clusters depicted here are in different proximity relations, and have different statistics. Their shapes vary from spherical (F, G, K) through ellipsoidal (C, D) and ellipsoid-like (H, I, J) to rectangular (E) and irregular (A, B). Some (B, C, D, F and G) are denser than others (A, E, J and K), and some (H and I) have inhomogeneous data distribution. Several (C, F, G) are very small, and two (J and K) are overlapping. Such variations produce a challenging structure that may not be effectively dealt with by methods which, by design, best handle clusters of given characteristics (such as spherical clusters by K-means).

were identified in this image and associated with meaningful physical entities such as roof materials, vegetation, roads, etc. [18]. A selected subset of those clusters is listed in Table 1, for the eight spectral bands (feature space dimensions) which were used in the earlier clustering. First, the number of data points varies extremely, from a few hundred to nearly 50,000, across clusters. Secondly, the standard deviation for each of the clusters varies across the spectral bands, and varies differently for each cluster (anywhere from 2- to 5-fold), indicating all kinds of shapes except hyperspherical. To compare the volumes of the clusters, we assume that clusters are hyperrectangles with a size proportional to the standard deviation in the corresponding dimensions (bands). While this does not give the correct volume of the clusters it still provides an insight to the magnitude of the variation. By this rough comparison, the largest cluster (L), is about 2000 times larger than the smallest cluster (G). We also compute the density of the clusters by dividing the number of data points by the calculated volume. The densest cluster (G) is about 3500 times more dense than the sparsest cluster (L). For example, G has about 28,000 data points, a volume of 4×10^{15} and a density of 7463×10^{-15} whereas I, which has a similar number of data points (about 25,000), has a volume of 211×10^{15} and a density of 117×10^{-15} (1/70 of the density of G). In Sect. 3.2 we will also show envelopes of these clusters (i.e., plots of extreme values in each feature dimension), from which it can be seen that many of the clusters are overlapping. The widely varying numbers of data points, volumes and densities, as well as overlaps of the clusters make this Ocean City data set complicated, highly structured.

Table 1. Statistics indicating complicated cluster structure in a remote sensing spectral image of Ocean City, Maryland. From over twenty identified clusters, corresponding to real physical entities in this image, nine selected clusters are listed here.

Cluster Labels	No. data points	standard deviation band 1	band 2	band 3	band 4	band 5	band 6	band 7	band 8	volume (10^{15})	density (10^{-15})
A	8089	145	88	171	101	136	102	175	181	96	84
B	5496	186	141	118	90	122	92	145	146	65	84
E	14684	186	109	76	102	126	81	179	229	65	225
G	28468	122	73	84	65	85	60	107	145	4	7463
H	480	176	109	118	112	115	81	156	214	78	6
I	24719	271	107	174	189	157	86	119	138	211	117
L	13592	339	268	207	207	233	148	238	269	8589	2
O	20082	341	221	157	159	213	148	283	275	4610	4
R	48307	179	119	163	106	107	109	235	269	271	178
V	998	177	229	112	79	92	81	180	217	106	9
a	239	354	286	205	175	168	144	201	177	3132	0.1

Extraction of such complicated clusters, and rather precisely, is important in many of today's real problems. For example (as we show in Sect. 4.2) some of the smallest clusters in this image represent unique roofing materials, and many of the clusters with considerable overlap in their signatures map distinct man-made materials. All of these are important to detect and map accurately in an urban development survey (for which this kind of data may be acquired). We want to point out that this Ocean City image is not as complicated as some of the hyperspectral images we have been working with. One such image will also be analyzed in Sect. 4.3. In the rest of this paper we discuss our contributions to the learning and extraction of clusters from highly structured manifolds.

3.1 Measuring the Correctness of SOM Learning for Complicated High-Dimensional Manifolds

A prerequisite of faithful cluster identification from an SOM is an appropriate representation of the topology of the data space M by the ordering of the prototypes in the SOM lattice A. Ideally, the SOM should be free of topology violations, at least in the forward direction since the "twists" caused by forward violations can lead to incorrect clustering. Backward topology violations are not detrimental for cluster extraction because they manifest in disconnects (strong dissimilarities — high fences — and/or prototypes with empty receptive fields, as in Fig. 1) in the SOM, which helps locate clusters. Ideal topology preservation usually does not occur for real data. For noisy, complicated manifolds topology violations are common at all stages of the learning. Adding to the difficulty is the fact that one does not know when the learning is mature enough, or how much further a seemingly static map may still improve. The key is to quantify

the extent of various violations and be able to separate important ones from the inconsequential.

The Topographic Product [28] (TP) was the first measure that quantified the quality of topology preservation. A prototype based measure, the TP is computationally economical, but it penalizes (falsely detected) violations caused by nonlinearities in the manifold, due to improper interpretation of neighborhood by Euclidean metric. This drawback is remedied in the Topographic Function [27] (TF), where the Euclidean metric is replaced by the graph metric of the induced Delaunay triangulation. Assuming a high enough density of prototypes in the manifold, the induced Delaunay graph can be constructed, after [26], by finding the best matching unit (BMU) and the second best matching unit (second BMU) for each data vector, and expressing these "connections" in a binary adjacency matrix of the SOM prototypes. Two prototypes that are a pair of BMU and second BMU for any data vector are adjacent or *connected* in the induced Delaunay graph. (Equivalently, they are centroids of adjacent Voronoi cells). The lattice (maximum norm) distance of two prototypes in the SOM is their *connection length* or *folding length* [27]. The TF not only uses a better distance metric, but also shows the scope of forward violations, by computing the average number of connections that exist at folding lengths larger than k:

$$TF(k) = \frac{1}{N} \sum_{i \in A} \{\# \text{ of connections of unit } i \text{ with length} > k\}, \tag{3}$$

where i is the index of the neural unit in A, N is the total number of units. A large k indicates a global (long range, more serious) violation while a small k corresponds to a local disorder. Eq. (3) is also applicable to backward violations, where k is negative and represents the induced Delaunay graph distance, in the data space, between prototypes that are adjacent in the SOM. There are also measures that only utilize information on the data distribution. For example, [32] introduced a cumulative histogram to express the stability of the neighborhood relations in an SOM. It captures a statistical view of the neighborhood status of the system and compares it with an unordered map. The more dissimilar they are, the more reliable the mapping. Another measure, the Topographic Error [29], expresses the extent of topology violation as a percentage of data points that contribute to violating connections. Extending these previous works, we enriched and further resolved the TF in the Weighted Differential Topographic Function (WDTF) [30].

$$WDTF(k) = \frac{1}{D}\{\# \text{ of data vectors inducing connections of length}= k\} \tag{4}$$

where D is the total number of data samples. The WDTF is a differential view of the violations at different folding lengths, in contrast to the integral view of the TF. It also adds new information by using the number of data samples that induce a given connection, as an *importance weighting*. By this weighting, the WDTF distinguishes the severity of violating connections: a long range but weak violating connection caused by a few noisy data vectors may be unimportant and safely ignored in the overall assessment of topological health, while a heavy violating connection warrants attention as a potential twist in the map.

Monitoring Violating Connections: TopoView and WDTF. Besides the topographic measure WDTF discussed above, we present another useful, interactive tool, TopoView, which allows to show violating connections on the SOM. This is more general than displaying SOM neighbors connected in data space, which is limited to 2- or 3-dimensional data. Different subsets of connections can also be selected by thresholding the connection strength and/or the connection length, which helps filter out noise, outliers and unimportant (weak) violations, and thereby more clearly see the relevant characteristics of the topology. For complex and high-dimensional data, it is especially effective to use both tools together. The WDTF provides a summary of the severity of violations at each folding length while TopoView provides localization of the violations in the SOM, for selected severity levels.

We illustrate the use of TopoView and the WDTF on a synthetic 4-class Gaussian data set, in Fig. 3. The data set was generated by using four Gaussian distributions with mean=0, standard deviation=1, at four centers in 2-dimensional space. At 1K (1000) steps (Fig. 3, top row), the SOM appears twisted in the data space, especially in the upper right cluster, where a chain of SOM protoypes is arranged in the shape of a horseshoe. TopoView reflects this twisting by a set of connections along the right side of the SOM. From the WDTF, which shows violations up to length 6, we can see what is known from the connection statistics: the end units of this chain, and also some of the non-neighbored units in between, must be connected. However, the long range violations are relatively weak. The rest of the violating connections are more local, but stronger, with a connection length of 2 or 3. As the SOM evolves, the set of long range connections in the SOM disappear at 3K steps (middle row), which means the "horseshoe" took up a shape that better approximates the spherical cluster. Finally, TopoView shows the SOM free of violating connections at 100K steps (bottom row): the prototypes are well placed in the data space, and the WDTF vanishes.

We give a real demonstration of the use of these measures through a hyperspectral urban image, which represents complicated, highly structured data. This image, which we will call "RIT image", was synthetically generated, therefore it has ground truth for every pixel, allowing objective evaluation of analysis results on the 1-pixel scale. The image pixels are 210-dimensional feature vectors (reflectance spectra), which are the data vectors input to the SOM. The scene contains over 70 different material classes. Fig. 4 and 5 give an illustration, and Sect. 4.1 a detailed description, of this data set.

To monitor topology preservation we compare two snapshots taken during the learning of the RIT image, at 500K and 3M (3000000) steps, respectively. Fig. 6 shows that the quality of the topology preservation improves from 500K to 3M steps. The number of short-range violations considerably decreases, and a decrease is generally showing at larger folding lengths, with some exceptions (such as at $k = 8$ and $k = 13$). From the TopoView representation in Fig. 7 one can follow which violations disappear between the two snapshots. Since TopoView shows the individual violating connections the thick cloud of connections can be

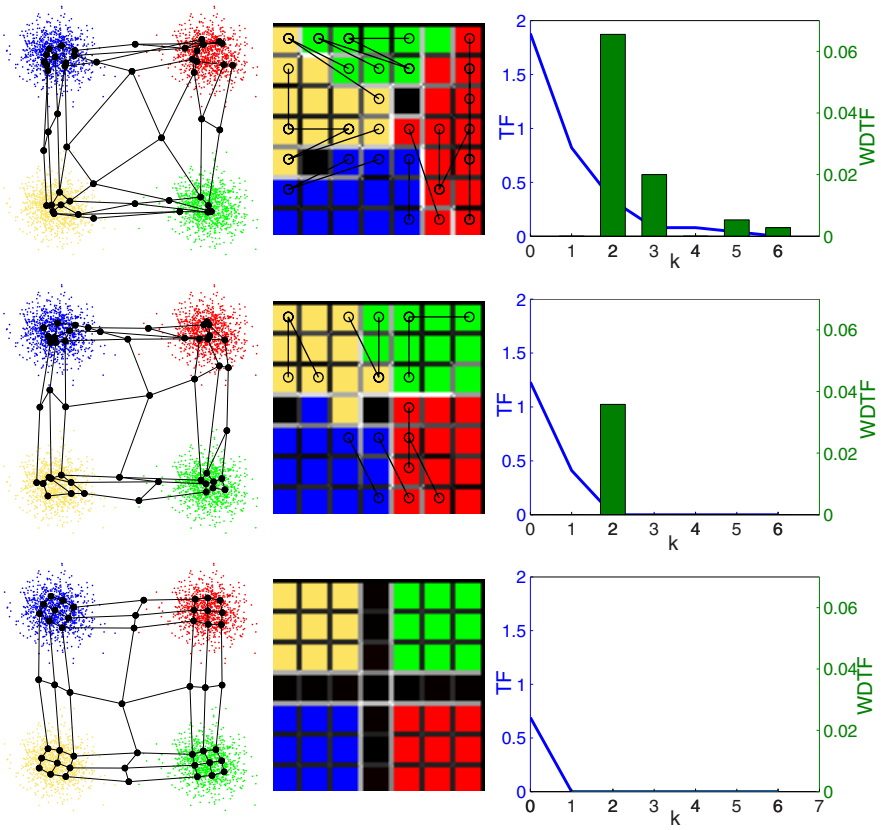

Fig. 3. The evolution of the SOM as it learns the synthetic 2-dimensional 4-class Gaussian data. Three snapshots are shown at 1K, 3K and 100K steps, from top to bottom. **Left**: The SOM prototypes (black dots) in the data space space, with SOM neighbors connected. Data samples are color coded according to their class memberships. **Center**: All violating connections shown as black lines, over the SOM, with the class labels (colors) and the mU-matrix also superimposed. **Right**: The TFs (blue lines) and the WDTFs (green bars).

obscuring. However, even with all connections shown (in the top row of Fig. 7), one can see that, for example, the lower left corner of the SOM became completely violation free at 3M steps. To give an "importance-weighted" view of the same we can apply tresholding by connection strength (the weighting used by WDTF). This eliminates unimportant connections and clears the view for analysis of those violations that may significantly contribute to cluster confusion. Two examples for possible thresholdings are given in Fig. 7. In both cases a decrease in confusion (relative to the mU-matrix fences) can be seen. Thresholding by connection length can make the cut between global and local range violations. This threshold depends on the data statistics and is automatically computed, based on the following argument, from [34]: if a prototype w_i has m Voronoi neighbors in data space

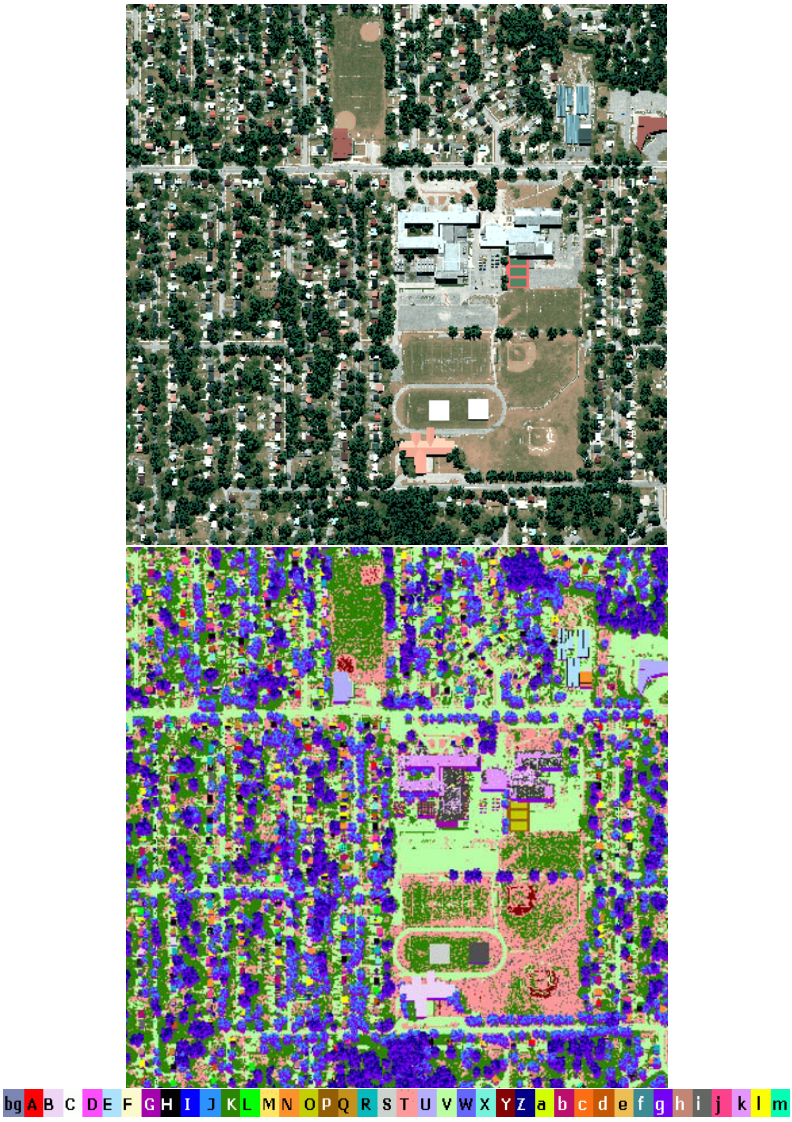

Fig. 4. Top: A color composite of 3 selected spectral bands of the 400 x 400 pixel, 210-band synthetic hyperspectral RIT image. **Bottom:** Partial cluster map, extracted from the self-organizing map in Fig. 5, using the modified U-matrix visualization [33]. It shows 39 cover types with unique colors keyed in the color wedge. Some additional unique materials (such as roofs of houses, showing in black) have no labels assigned for reasons of color limitations. Besides the obvious vegetation (trees and grasses, clusters I, J, K, Z, g, Y), the approximately 70 different surface cover types in this image include mixed dirt/grass (T), a large number of roof materials (A, B, C, F, M, N, Q, R, U, X, a, b, d, h, i, j, k, l, m), pavings (V, roads and parking lots), tennis courts (O, P), several types of car paints (W, c, M, f, e), and glass (windshield of cars and roof, E).

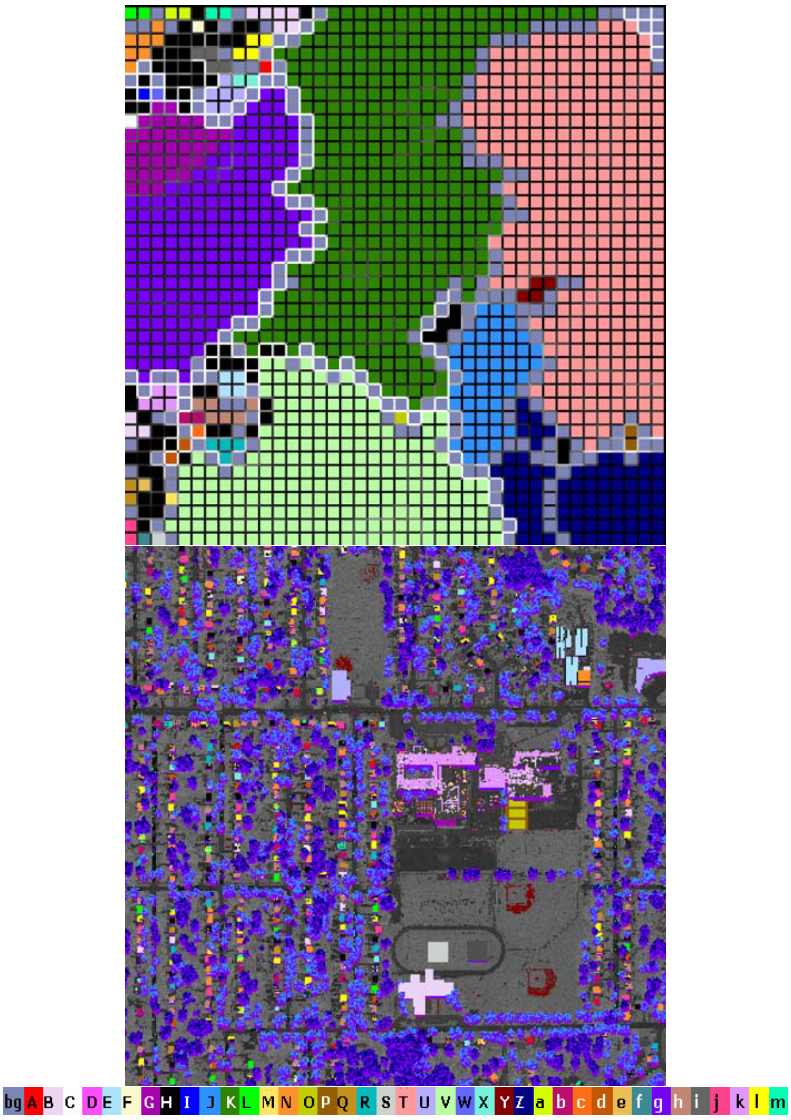

Fig. 5. Top: The SOM with discovered clusters color-coded and the mU-matrix superimposed. Interpretation of these clusters is given in Fig. 4. Medium grey cells (the color of the background, "bg"), appearing mostly along cluster boundaries are SOM prototypes with no data points mapped to them. Some prototypes — shown as black cells, which have data mapped to them — were left unclustered, because of color limitation. **Bottom:** The same cluster map as in Fig. 4, bottom, with the large background clusters (grass, paved roads and lots) removed to provide better contrast for the many different roof materials, and other small unique spectral clusters such as tennis courts. About twenty of these clusters are roof types. Spectral plots showing excellent match of the spectral characteristics of the extracted clusters with true classes are in [33].

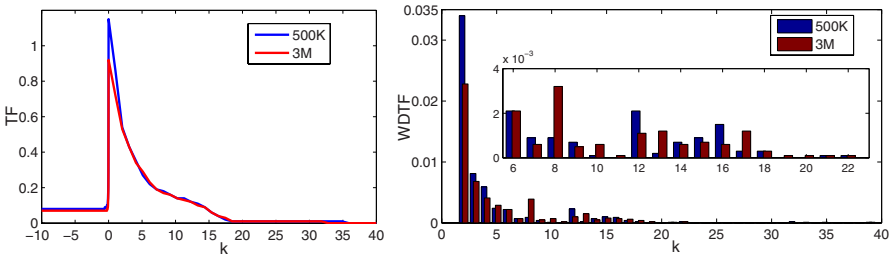

Fig. 6. Comparison of TFs (**left**) and WDTFs (**right**) after 500K and 3M SOM learning steps on the RIT hyperspectral image. While the integral measure TF shows a general decrease of violations at shorter folding lengths, the WDTF also indicates the fluctuations of violations across folding lengths.

then a topology preserving arrangement in the SOM lattice for these m neighbors is a placement into the "tightest" SOM neighborhood. This means that the 8 closest Voronoi neighbors should occupy the 8 immediate lattice neighbors of w_i in a square SOM neighborhood, Voronoi neighbors 9 – 24 wrap around this first tier of immediate SOM neighborhood, and so on. The radius of the SOM neighborhood that accomodates all Voronoi neighbors in this tightest fashion yields the folding length k within which the violations can be considered local. In the case of the RIT data, the maximum number of connected neighbors is 21 at 500K steps and 19 at 3M steps. These can fit into a 2-tier (8+16) square SOM neighborhood. Therefore, global violations are those with $k \geq 3$, shown in Fig. 7, bottom row. TopoView can also show inter-cluster or intra-cluster connections separately if a clustering is provided, and thus aid in the verification of clustering. This will be shown in Sect. 4.3.

3.2 Cluster Extraction

Cluster extraction from SOMs is accomplished through the clustering of the learned prototypes. Various approaches can be used, either based solely on the similarities of the prototypes or by taking into account both the prototype similarities and their neighborhood relations in the SOM grid. The latter is typically done interactively from visualization of the SOM and is generally more successful in extracting relevant detail than automated clustering of the prototypes with currently available methods.

SOM knowledge representations — what information is quantified and how — are key to the quality of cluster capture from visualizations. The widely used U-matrix [35] displays the weight distances of the SOM neighbor prototypes, averaged over the neighbors and coloring the grid cell of the current prototype to a grey level proportional to this average distance. The U-matrix and its variants (e.g., [36], [37], [38]) are most effective when relatively large SOM grid accomodates small data sets with a low number of clusters because the averaging can obscure very small clusters and sharp boundaries in a tightly packed

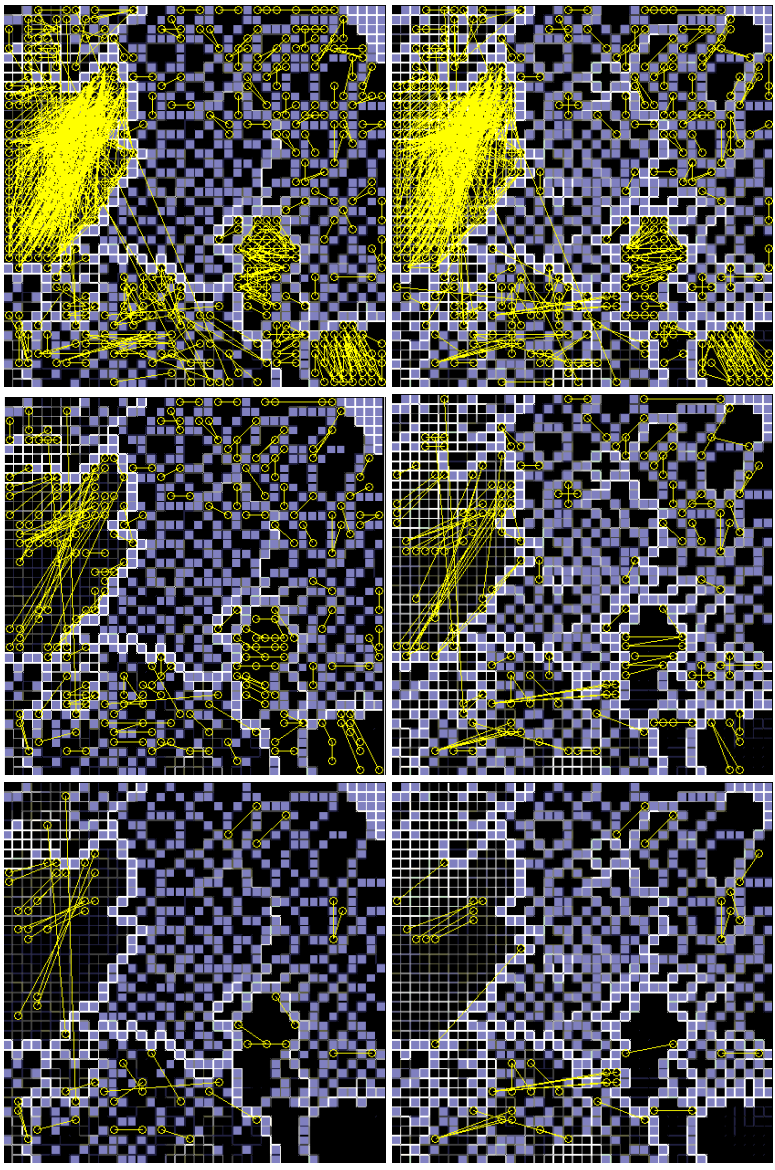

Fig. 7. TopoView visualization of violating connections (yellow lines) with different thresholdings, on the SOM of the RIT data at 500K steps (left column) and at 3M steps (right column). In the underlying SOM, the mU-matrix is superimposed (white fences). Medium grey cells indicate empty prototypes. **Top:** all violating connections are drawn. **Center:** violating connections with strength greater than the mean strength of all violating connections. **Bottom:** global violating connections ($k \geq 3$) with connection strength greater than the mean strength of the fourth strongest connections of all prototypes. This choice of threshold, proposed in [34], is described in Sect. 3.1.

SOM grid. Approaches such as [39] and gravitational methods (e.g., Adaptive Coordinates [37]) visualize distances between the weights in innovative ways that greatly help manual cluster extraction. Automated color assignments also help qualitative exploration of the cluster structure [40], [41], [42]. We point the reader to [37], [43], [34] for review. Visualization of the size of the prototype receptive fields (e.g., [38], [44]) is among the earliest tools. Visualization of samples that are adjacent in data space but map to different SOM prototypes [45] is a richer representation than the previous ones since it makes use of the data topology. However, in case of a large number of data points, adjacent samples mapped to different prototypes are only the ones at the boundaries of the Voronoi polyhedra, thus this visualization still leaves a lot of the topological information untapped. More of the data topology is utilized by [46] and [47], however, the visualization is in the data space and therefore limited to up to 3 dimensions.

We added to this arsenal the *modified U-matrix* (mU-matrix), and the *connectivity matrix*, CONN, and their visualizations. They are especially useful for large, highly structured data sets mapped to not very large SOMs. (The size of the SOM is a sensitive issue with high-dimensional data as the computational burden increases non-linearly with input dimension. One wants to use a large enough SOM to allow resolution of the many clusters potentially present in the data set, but not exceedingly larger than that.) The mU-matrix is a higher resolution version of the U-matrix in that it displays the weight distances to all neighbors separately, on the border of the grid cells including the diagonals, as shown in Fig. 1. Combined with the representation of the receptive field sizes (red intesities of the grid cells in Fig. 1, left), it conveys the same knowledge as in [39]. While the sense of distance in the mU-matrix is not as expressive as the "carved away" grid cells in [39] the mU-matrix leaves room for additional information to be layered. Examples of that are [48] and [49] where known labels of individual data objects were displayed in the grid cells thereby showing, in addition to the density, the distribution of the known classes within receptive fields. The mU-matrix is advantageous for the detection of very small clusters, as for example, in Fig. 5, where many small clusters are represented by just a few (even single) protoypes in the upper and lower left corners.

The CONN knowledge representation was first proposed in [50], developed for visualization in [51], and is presented in detail in [34]. It is an extension of the induced Delaunay triangulation, by assignment of weights to the edges of the graph. An edge connecting two prototypes is weighted by the number of data samples for which these prototypes are a BMU and second BMU pair. This weighting is motivated by the unisotropic distribution of the data points within the Voroni cells, as explained in Fig. 8 on the "Clown" data created by [52].

The edges of the weighted Delaunay graph can be described by

$$CONN(i,j) = |RF_{ij}| + |RF_{ji}| \tag{5}$$

where RF_{ij} is that section of the receptive field of w_i where w_j is the second BMU, and $|RF_{ij}|$ is the number of data vectors in RF_{ij}. Obviously, $|RF_i| =$

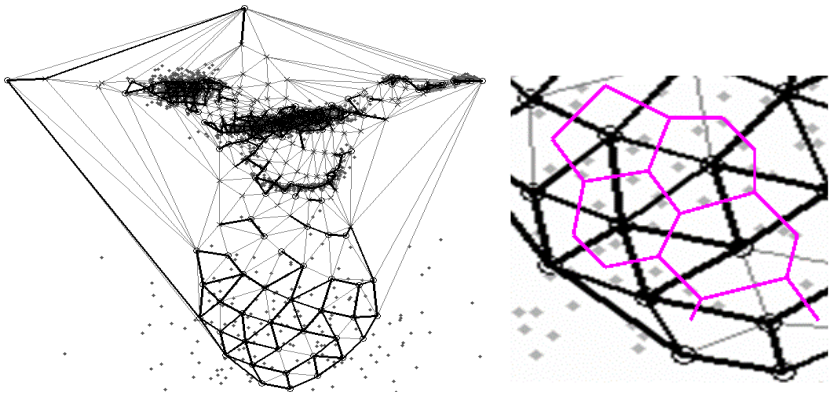

Fig. 8. Left: Delaunay triangulation (thin lines) and induced Delaunay triangulation (thick lines) for the 2-d "Clown" data set, using the SOM prototypes created by [52]. The "Clown" is indicated by the annotations in Fig. 9. We omit annotations here so as not to obscure details. The small dots represent data points. Prototypes with non-empty receptive fields are denoted by circles, prototypes with empty receptive fields are shown by 'x'. The induced Delaunay triangulation exposes discontinuities in the data manifold, for example, the separations between the eyes, the nose and the mouth, while the Delaunay triangulation does not highlight them. **Right:** Magnified detail from the lower left part of the "Clown". Data points in the Voronoi cells, superimposed in pink, exhibit an unisotropic distribution, indicating variable local data densities in the directions of the Voronoi neighbors.

$\sum_{j=1}^{N} |RF_{ij}|$ because $RF_i = \cup_{j=1}^{N} RF_{ij}$. CONN thus shows how the data is distributed within the receptive fields with respect to neighbor prototypes. This provides a finer density distribution than other existing density representations which show the distribution only on the receptive field level. $CONN(i, j)$, the connectivity strength, defines a similarity measure of two prototypes w_i and w_j.

Visualization of this weighted graph, CONNvis, is produced by connecting prototypes with edges whose widths are proportional to their weights. The line width gives a sense of the *global importance* of each connection because it allows to see its strength in comparison to all other connections. A *ranking* of the connectivity strengths of w_i reveals the most-to-least dense regions local to w_i in data space. This is coded by line colors, red, blue, green, yellow and dark to light gray levels, in descending order. The ranking gives the relative contribution of each neighbor independent of the size of w_i's receptive field, thus the line colors express the *local importance* of the connections. The line width and the line color together produce a view of the connectedness of the manifold, on both global and local scales. This is shown in Fig. 9 for the "Clown" data. We use this 2-dimensional data set because it has an interesting cluster structure, and because we are able to show the information represented by CONN both in data space and on the SOM, thus we can illustrate how CONNvis shows data structure on the SOM regardless of the data dimension. Compared to Fig. 8, left, all connections remain, but now the connection strengths emphasize strongly and

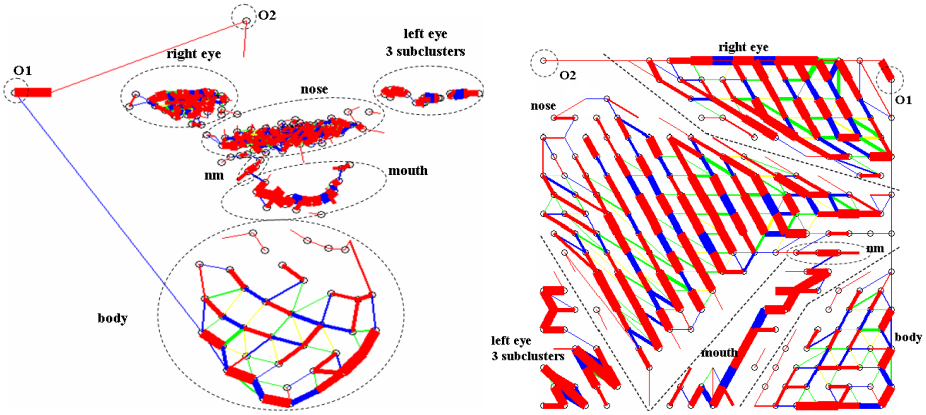

Fig. 9. Left: The connectivity matrix CONN (weighted Delaunay triangulation) shown on the 2-d "Clown" data set, using the SOM prototypes created by [52]. Parts of the "Clown" are explained by the annotations. O1 and O2 are outliers. The lack of a circle symbol indicates an empty prototype. Line widths are proportional to the weight of the edge, $CONN(i, j)$, expressing its global importance among all connections. Colors show the ranking, the local importance of the connections to the Voronoi neighbors. This is redundant with the line widths but because we bin the line widths to help the human eye distinguish grades of connection strengths, color coding the ranking restores some of the information lost through the binning. The ranking is not symmetric, *i.e.*, if the rank of w_j for w_i is r, and the rank of w_i for w_j is s, r is not necessarily equal to s. The connections are drawn in the order of lowest to highest rank so a higher-ranking connection will overlay a lower-ranking one. Details of subcluster structures are visibly improved compared to Fig. 8. **Right:** The same CONN matrix draped over the SOM.

poorly connected (high and low density) regions. Clusters not obvious in Fig. 8 and not visible in the U-matrix of the Clown in [52] such as the three subclusters in the left eye, clearly emerge here.

One significant merit of CONNvis is that it shows forward topology violations in a weighted manner, giving a strong visual impression of the densest textures in the data. CONNvis is somewhat limited in resolving many connections because the weighting (line width) can obscure finer lines in a busy CONNvis. Complementarily, TopoView can show many connections simultaneously, in selected ranges of the connection strength. Alternative use of these two visualizations, which render the same knowledge, can be quite powerful. Both CONNvis and TopoView also show backward violations through unconnected SOM neighbors. These indicate discontinuities in the manifold and thus immediately outline major partitions in the data. The mU-matrix has capacity to indicate backward topology violations through corridors of prototypes with empty receptive fields, such as in Fig. 1, and through high fences but not as clearly as CONNvis or TopoView, and it cannot show forward violations.

Cluster Extraction with CONNvis. Interactive clustering with the help of CONNvis can be done by evaluation and pruning of the connections. Unconnected

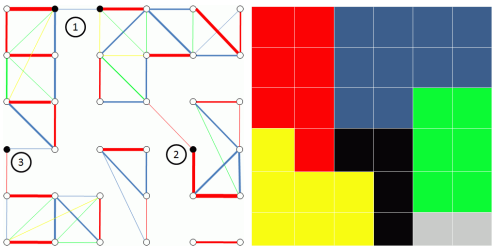

Fig. 10. Left: Illustration of interactive clustering with CONNvis. Strongly connected groups of prototypes (coarse clusters) visually emerge through the lack of connections or weak connections across some prototypes (annotated by black dots), easily recognized by a human analyst. The connections of these straddling prototypes present one of three situations for which the following rules are applied: 1: If a prototype has different number of connections to each coarse cluster, cut the smaller set of connections; 2: If a prototype has the same number of connections to each cluster but with different connectivity strengths, cut the weaker set of connections; 3: If a prototype has the same number of connections to each cluster, with the same strengths but different rankings, cut the lower ranking ones. **Right:** The identified clusters.

or weakly connected prototypes in the unpruned CONN often already outline "coarse" clusters, i.e., densely connected areas in the SOM which have considerably fewer connections to one another. The weakly connected prototypes at the boundaries of the coarse clusters are easily recognized by the human operator (black dots in Fig. 10). The corresponding weak connections are then evaluated as described in Fig. 10, to find and severe the "weakest link" for proper cluster separation.

For complicated cases, where the number of data points is huge, and the data are noisy, prototypes can have a large number of connections (neighbors in data space), and also a relatively large number of connections across coarse clusters. This creates a busy CONNvis, and requires considerable work when the number of clusters in the data is large, but the procedure is exactly the same as in simple cases. It has been used to produce some of the results shown here, and elsewhere [33]. At this time of writing we are collecting experiences from interactive clustering with CONNvis, which we expect to turn into an automatic procedure ultimately.

4 Case Studies with Higly Structured Real Data Sets

4.1 Real Data Sets

The Ocean City Multispectral Image. was obtained by a Daedalus AADS-1260 multispectral scanner. The image comprises 512×512 pixels with an average spatial resolution of 1.5 m/pixel [31]. Each pixel is an 8-dimensional feature vector (spectrum) of measured radiance values at the set of wavelengths in the 0.38–1.1μm and 11–14 μm windows that remained after preprocessing [53].

Ocean City, along the Maryland coast, consists of rows of closely spaced buildings separated by parallel roads and water canals. The spatial layout of different surface units is shown in Fig. 11, left, through an earlier supervised class map [53], which we consider a benchmark since it was verified through aerial photographs and field knowledge [31,54]. Ocean (blue, I) surrounds the city, ending in small bays (medium blue, J, at the top center and bottom center of the scene) which are surrounded by coastal marshlands (brown, P; ocher, Q). Shallow water canals (turquoise, R) separate the double rows of houses, trending in roughly N-S direction in the left of the scene and E-W direction in the right. Many houses have private docks for boats (flesh-colored pink, T) and as a consequence, dirty water at such locations (black, H). Paved roads (magenta, G) with reflective paint in the center (light blue, E) and houses with various roof materials (A, B, C, D, E, V) show as different classes. Typical vegetation types around buildings are healthy lawn, trees and bushes (K, L), yellowish lawn (split-pea green, O) and dry grass (orange, N).

The RIT Hyperspectral Image. briefly introduced in Sect. 3.1 and in Fig. 4 and 5, was synthetically generated through rigorous radiative transfer modelling called the DIRSIG procedure at the Rochester Institute of Technology [55,56]

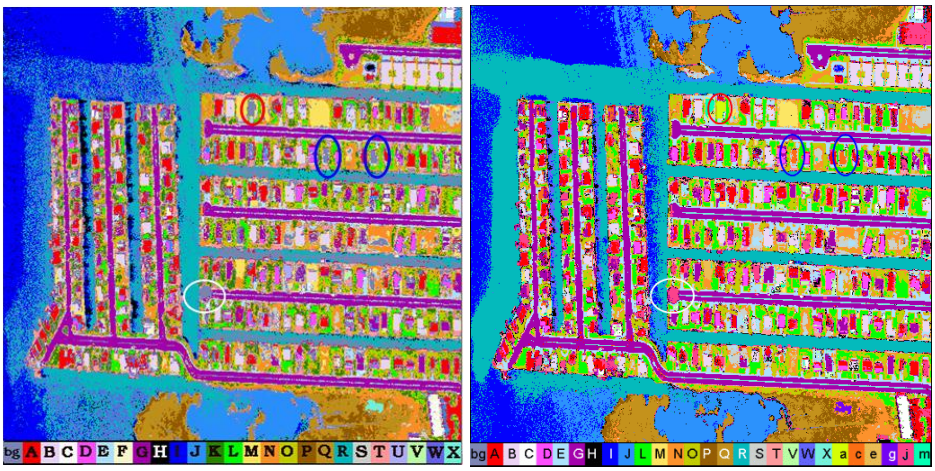

Fig. 11. Left: Supervised classification of the Ocean City image, mapping 24 known cover types. Red, white and blue ovals show unclassified shapes of buildings and a circle at the end of a road (the color of the background, bg). **Right:** Clusters identified interactively from CONNvis visualization of a SOM of the Ocean City image. The agreement between the cluster map and supervised class map is very good. The unclassified gray spots (in red, white and blue ovals on the left) are now filled exactly, and with colors (a, c, j) different from the 24 colors of the supervised classes. These new clusters only occur at the locations shown, indicating the discovery of rare roof types and road materials, which were not used for the training of the supervised classifier. The spectral signatures of the newly discovered clusters, as well as those of two new subclusters (e, m) are distinct from the rest.

(hence the name RIT). Owing to its simulated nature, this image has ground truth for every pixel, allowing objective evaluation of analysis results on the 1-pixel scale. The realism of the RIT image is quite amazing, in both spectral and spatial respect. Its characteristics are close to that of an AVIRIS image. (AVIRIS is the Airborne Visible Near-Infrared Imaging Spectrometer of NASA/JPL [57,58], to date the most established, extremely well calibrated and understood airborne hyperspectral imager.) The scene comprises 400 x 400 pixels in 210 image bands in the 0.38 to 2.4 μm visible-near-infrared spectral window. The spatial resolution is 2 m/pixel. Realistic noise and illumination geometry is part of the simulation. The visual appearance of a natural color composite made from three selected image bands in the red, green, and blue wavelength regions, Fig. 4, top, is virtually indistinguishable from an image of a real scene. It contains over 70 different classes of surface materials, widely varying in their statistical properties in 210-dimensional data space. The materials include vegetation (tree and grass species), about two dozens of various roof shingles, a similar number of sidings and various paving and building materials (bricks of different brands and colors, stained woods, vinyl and several types of painted metals), and car paints. Many of these materials are pointed out in Fig. 4. This image was clustered from mU-matrix visualization in [33], as shown in Fig. 4 and 5.

4.2 Clustering of the Ocean City Multispectral Image

We show clustering with CONNvis on an image of Ocean City, described in Sect. 4.1. Fig. 11 compares an earlier benchmark supervised classification (24 classes, [18]) with an SOM clustering obtained through CONNvis as described in Fig. 10. Fig. 12 shows the extracted 27 clusters on the SOM and an enlarged part of the CONNvis for details of cluster separations including several small clusters. The two thematic maps in Fig. 11 have a strong similarity, which suggests that the clustering found all supervised classes. In addition, it discovered several new ones (a, c, j), which were not known at the time of the supervised classification. From comparison with aerial photographs these appear to be roofs (clusters "a" and "c" in red and blue ovals), and a different surface paving (cluster "j", in white oval) at the end of one road. These and other new units (e, m, subclusters of supervised class M) are distinct enough spectrally to justify separate clusters, as seen in Fig. 13. Another important improvement in the CONNvis clustering is that it assigned labels to many more pixels than the earlier supervised classification, which manifests in more green (vegetation) and turquoise (ocean water) pixels. This is not only because of the discovery of new material units (which are very small) but mostly because the CONNvis view helps quite precise delineation of the cluster boundaries (as shown in Fig. 12).

Comparison with the popular ISODATA clustering is interesting. ISODATA clustering, done in ENVI (ITT Industries, Inc., http://www.ittvis.com/index.asp), was graciously provided by Prof. Bea Csathó of the University of Buffalo [54]. The ISODATA is a refined k-means clustering [59] that has the flexibility to iteratively come up with an optimum number of clusters capped at

a user specified maximum. Using the default parameters values in ENVI for a maximum of five iterations, 10 clusters, shown in Fig. 14, left, were produced when up to 10 cluster centers were allowed. The 18 clusters in Fig. 14, right, resulted for a maximum of 20 (and also for 30) clusters. Experiments allowing more iterations and more clusters to be merged produced no visible change. To help visual comparison with the SOM maps, we tried to recolor the randomly assigned ISODATA colors to those in the cluster map in Fig. 11, by assigning to each ISODATA cluster the color of that SOM cluster which is most frequent in the given ISODATA cluster. (This obviously has limits since clusters formed by two different algorithms are not necessarily the same. For the same reason the color wedges and labels of each cluster map are different.) For example, the ISODATA cluster G was assigned the color of SOM cluster G (road, concrete) since the road surfaces dominate that ISODATA cluster. This recoloring immediately shows that in the 10-cluster case (Fig 14, left) this ISODATA cluster also comprises several roof types (SOM clusters B, C, D, E, F, U and V) that are spectrally distinct and resolved in the SOM map. ISODATA formed superclusters of the 27 clusters in the SOM map. Similar supergroups can be seen for the vegetation.

The 18-cluster case (Fig 14, right) is more complicated, but ISODATA still formed recognizable superclusters. The correspondence between water bodies is obvious. SOM clusters G (road) and B now have one-to-one match with the same labels in the ISODATA map, but there is also confusion among clusters. For example, B (orchid color, concrete roof), in the ISODATA map also includes SOM clusters E (the divider paint and a roof type, light blue) and M, a bare lot (yellow, at the top row of houses), in spite that E and M are distinct spectrally.

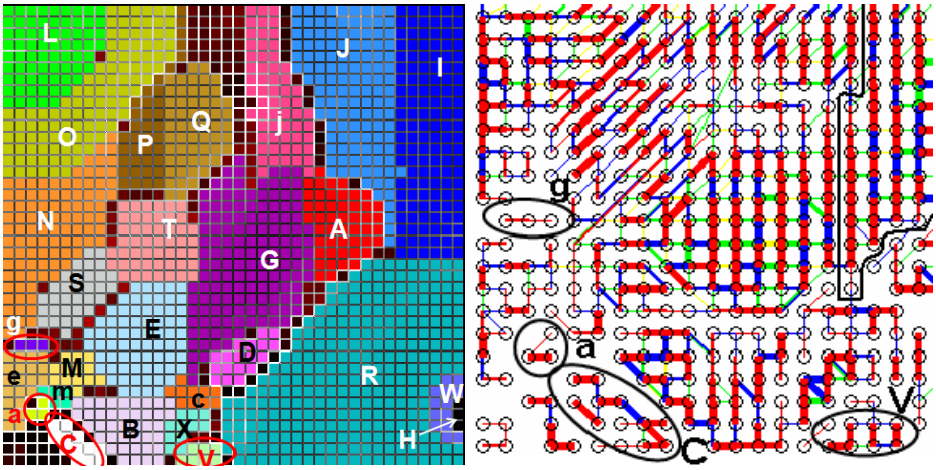

Fig. 12. Cluster extraction for the Ocean City data from CONNvis. **Left:** The extracted clusters in the SOM shown with the same color codes as in Fig. 11, right. **Right:** CONNvis of the bottom left quarter of the SOM to illuminate the representation of cluster boundaries. The small clusters C, g, V and a are clearly separated.

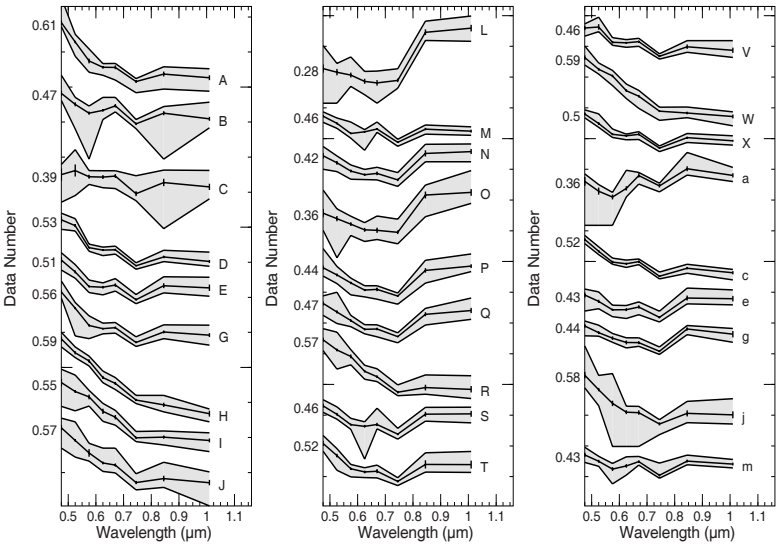

Fig. 13. Spectral statistics of the 27 SOM / CONNvis clusters of the Ocean City image (Fig. 11, right). Mean spectra with standard deviations (vertical bars) and the envelope of each cluster (shaded area) are displayed, vertically offset for viewing convenience. The number at left of each curve indicates the DN value in the first channel. Most of the large clusters (A, B, E, G, I, J, L, N, O, P, Q, R, S, T, j) are tight, suggesting clean delineation of boundaries.

The spectral plots in Fig. 13 and 15 also give partial indication of this confusion, relative to SOM clustering. It is obvious that in both cases the spectral clusters have overlaps (as expected). There is, however, sufficient discriminating information in the non-overlapping bands that the SOM / CONNvis clustering was able to utilize. Most of the 27 CONNvis clusters are reasonably tight (their envelope follows the mean, and standard deviations are small). 6–8 (about one fourth) of the clusters appear to have outliers indicated by loose envelopes but still small standard deviations. In comparison, half of the 18 ISODATA clusters have loose envelopes. More interestingly, most of the large clusters (see listing in the caption of Fig. 13) are tight in the CONNvis plots whereas most of the large clusters (listed in the caption of Fig. 15) are loose in the ISODATA case. This suggests that the boundary delineations by CONNvis, which are in agreement with the benchmark classification, are cleaner. ISODATA forces spherical clusters, whose boundaries may significantly differ from the natural cluster boundaries.

As shown in [53] for a 196-band hyperspectral image of another part of Ocean City containing 30 verified clusters, with increased complexity of the data (when also a larger number of clusters were allowed for ISODATA) ISODATA's confusion of the true clusters greatly increased compared to what we show here for the multispectral case.

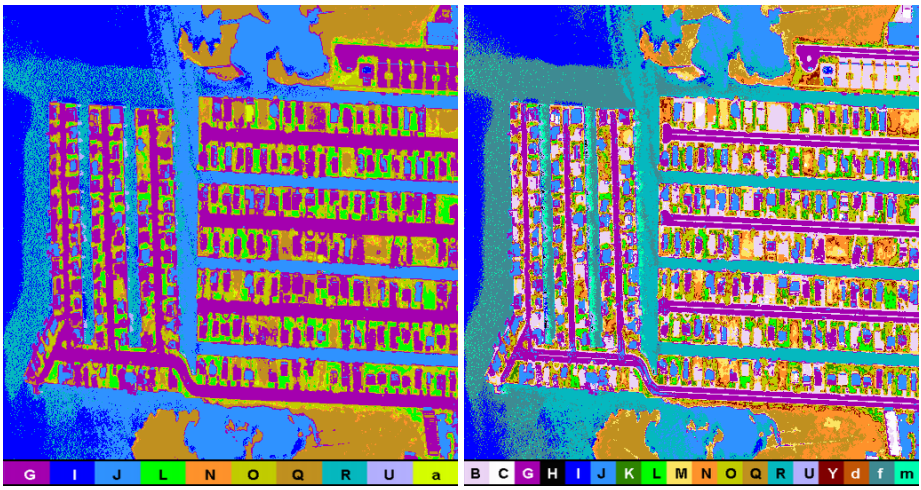

Fig. 14. ISODATA clusterings of the Ocean City image. **Left:** 10 clusters resulting from allowing 5 to 10 clusters. Examination of details reveals that the ISODATA clustering represents quite clean cut superclusters of the SOM clusters, as discussed in the text. **Right:** 18 clusters resulting from allowing 10 to 20 clusters. The ISODATA clusters still form supergroups of the SOM clusters, but some confusion also occurs. The spectral plots in Fig. 13 and 15 provide an insight to, and comparison of cluster separations.

4.3 Clusterings of the RIT Synthetic Hyperspectral Image

The 210-band RIT image, described in Sect. 4.1 and in Fig. 4, presents a case of extreme variations among a large number of clusters. This diversity is reflected in Fig. 5 where 39 extracted clusters are overlain on the SOM of this data set, and the spatial distribution of the clusters in the image is also shown. The number of data points in a cluster varies from 1 to nearly 40,000. Many clusters (twenty some different roof types of single houses) have only 200–400 pixels, and several makes of cars (clusters c, f, W, e, noticable mostly in the parking lot in the center of the scene) are represented by less than 10 image pixels. These very small clusters each occupy 1–2 prototypes at the upper and lower left corners of the SOM, along with a few groups of 4–6 prototypes (for example U (lilac), B (orchid), k (medium purple), or E (light blue)), which map larger buildings, very apparent in the scene. The spectra of these cover types exhibit a wide range of similarities. For example, the paving (cluster V, light green), and the grass (K, pure green) are very different, indicated by the strong mU-matrix fences; while a subset of the asphalt shingle roofs have quite subtle yet consistent differences [33]. (As explained in Sect. 3.1 color limitations restricted us to show only about half of the more than 70 clusters. We also used a common color code for several vegetation types, dark blue clusters in the lower right of the SOM, to make color variations more effective for the small clusters which are of greater interest.)

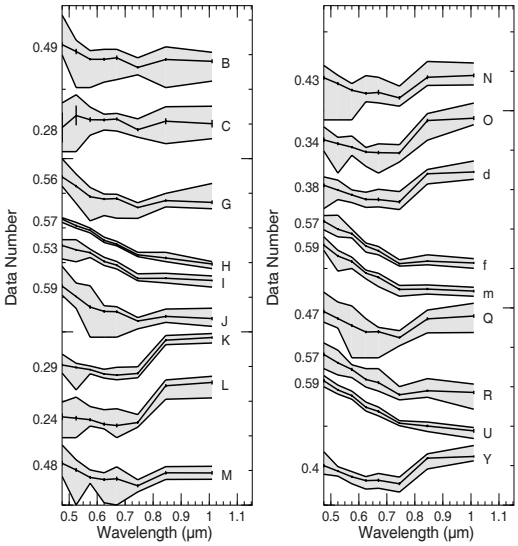

Fig. 15. Spectral statistics of the 18 ISODATA clusters of the Ocean City image (Fig. 14, right). Mean spectra with standard deviations (vertical bars) and the envelope of each cluster (shaded area) are displayed, vertically offset for viewing convenience. The number at left of each curve indicates the DN value in the first channel. Most of the large clusters (B, G, I, J, K, L, M, N, O. Q, R) appear loose.

The details of this clustering from a mU-matrix representation, including descriptions of the surface materials (cover types), spectral characterization showing similarities and differences, matches with the known true classes, and demonstration of discovery of various cars (tiny clusters), are published in [33]. Since we can capture more details with either mU-matrix or CONNvis than with ISODATA (as shown in Sect. 4.2 and in [53]), here we want to examine the relative merits of mU-matrix and CONNvis representations for extraction of clusters from this complicated data set.

We show two clusterings from two SOMs in Fig. 16, which were learned separately but with the same parameters and both to 3M steps. Consequently they are very similar with some minor differences, thus we can make comparative observations between these two clusterings. The top row of Fig. 16 presents the one from mU-matrix visualization in [33], the bottom row shows one that resulted from CONNvis clustering. (Owing to random assignments the same clusters have different colors in the two maps, but the very similar layout helps relate them visually. The cluster labels we use in this section refer to the key in the color wedge in Fig. 5.) The mU-matrix fences are also superimposed. In addition, empty prototypes are shown as medium grey cells (mostly at the boundaries of major clusters). Empty prototypes can be overlain by cluster labels (colors) to produce a homogeneous look of a cluster such as in the case of most of the large clusters here. We removed the color label of cluster V (light green in the top

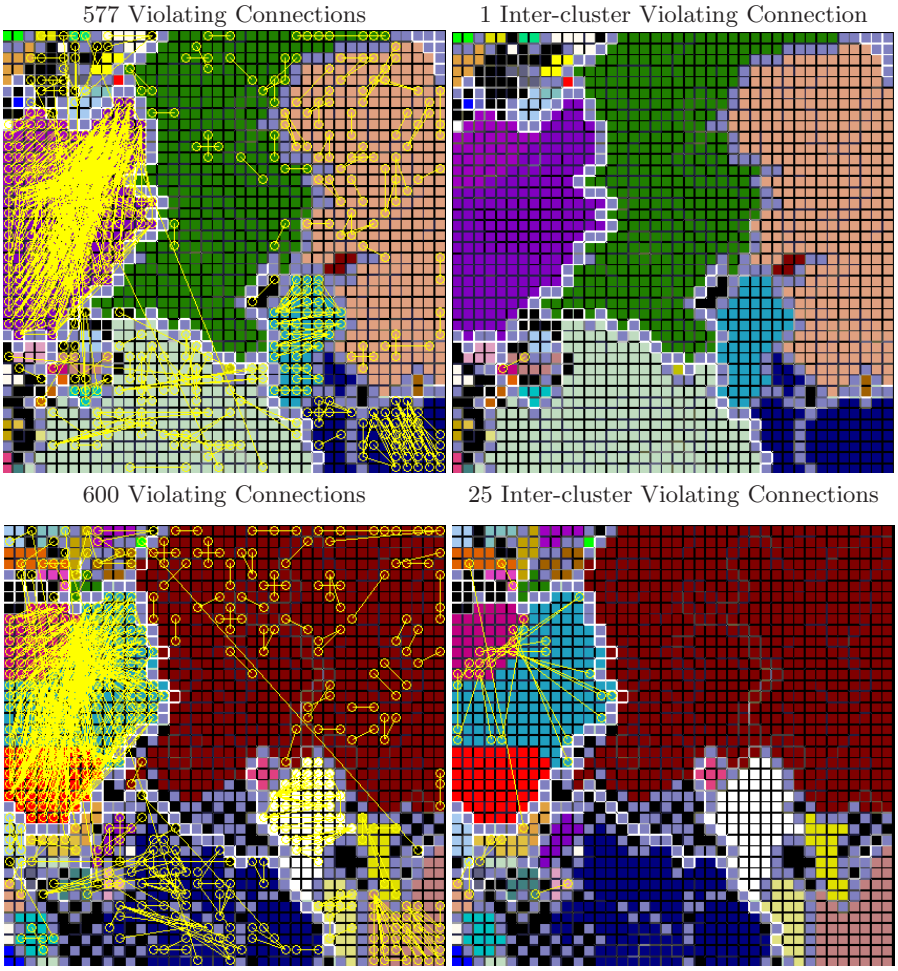

Fig. 16. Comparison of two clusterings of the RIT data. **Top** row: clusters extracted from mU-matrix knowledge, and color coded as in Fig. 5. **Bottom** row: clusters captured from CONNvis. Color coding of clusters is different from the top row because of random label assignments. The mU-matrix fences are superimposed on both. Medium grey cells are empty prototypes, whereas black cells are prototypes left unclustered for reasons of color limitation. **Left**: TopoView visualization (yellow lines) of all violating connections are overlain. **Right**: inter-cluster violating connections are shown.

row) and from the lower part of the large cluster K (pure green) in the CONNvis clustering (bottom row) to show the underlying scattered empty prototypes. We overlaid, on top of the extracted clusters, the TopoView showing all violating connections as yellow lines on the left, and inter-cluster violating connections on the right (i.e., violating connections which have two ends in the same cluster, or either end is an empty or unlabelled prototype were omitted). TopoView

confirms that the two SOMs are very similarly organized. One significant detail is that in both maps there are few inter-cluster violating connections going into or out of the small clusters in the upper and lower left corners.

The first impression is that the two clusterings are very similar. The most striking difference involves the large clusters K (pure green, grass) and T (flesh color, grass/dirt mix) at the center and right of the SOM at top, which were extracted as one cluster (largest, rust color area) in the SOM at the bottom. At top, a corridor of empty protoypes is clearly visible between the clusters K and T. In the mU-matrix representation, this corridor is accompanied by a double fence of consistent height, indicating a uniform difference between K and T along this corridor. (The absolute height of this fence is lower than that of the other, more prominent ones in this figure, but it can be seen unmistakably by interactive setting of the visualization range for fence values.) Similar corridor exists under the large rust color cluster in the bottom. However, many other empty prototypes are also present in both SOMs under these clusters, mostly in the checkered layout as seen at the unclustered part of the bottom SOM (or in Fig. 7). In the CONNvis these are all unconnected (no connection goes outside of the manifold, not shown here), therefore the empty cells could outline cluster boundaries the same way as the unconnected prototypes do in Fig. 12 for the Ocean City data. The difference is that in the case of Ocean City a corridor was cut by severing similarly weak connections in a contiguous area of non-empty prototypes, whereas here there are no connections to evaluate, the discontinuities have equal importance and therefore the same corridor does not emerge from the checkered pattern. As a result CONNvis sees these two units as a field of many small clusters. Since CONNvis lacks the distance information from which to notice that some of the discontinuities caused by these small clusters are "more discontinuous" than others, one cannot infer that there are two groups of small clusters in this field between which the (distance based) spectral dissimilarity is much greater than the dissimilarities within each of these two groups. The same scattered empty prototypes may be the cause of some small clusters not being apparent in the CONNvis, for example the tennis court, which is mapped to one prototype in the upper SOM (O, split pea color), between the light green cluster V and cluster K; or the running paths of the baseball field (cluster Y, rust color, wedged into cluster T in the top SOM). Under these circumstances CONNvis may not have the tool to distinguish those single-prototype clusters that are usually obvious from their "fenced off" appearence in a mU-matrix.

It is interesting to note that the relatively large cluster "g" (purple, at center left of the top SOM) is the most disorganized according to TopoView: it has the most violating connections. The reason is that this cluster is extremely noisy. In contrast, the largest clusters seem well organized with only spurious topology violations. However, the boundaries of all clusters including "g" were cleanly extracted, which is indicated by the lack of inter-cluster violations at top right. These connections were not thresholded by strength, the few showing in the bottom right SOM are mostly weak and unimportant.

TopoView reveals another important difference between the two clusterings: the splitting of the large cluster "g" into two (turqoise and red) by CONNvis. These two clusters are much more similar to one another in their spectral statistics (mean, standard deviation, not shown here), i.e., by distance based similarity, than the clusters K and T discussed above. Yet, CONNvis separates them by connectivity measure. In contrast to the clusters K and L, here the SOM has a contiguous field of non-empty prototypes, thus the relative connectivity strengths can be evaluated and cuts made as prescribed for CONNvis cluster extraction (Fig. 10). These two subclusters were not visible from the distance based (mU-matrix) representation, where the entire parent clsuter "g" appears to have a fairly uniform mesh of high fences.

Concluding from these discussions, the mU-matrix can be difficult to interpret where prototype distances may be very similar but relatively large within clusters with large variance. In contrast, CONNvis can be blind to distance based similarities. This susggests that alternating or using these two representations together would further increase the effectiveness of cluster extraction.

5 Conclusions and Outlook

We concentrate on issues related to cluster identification in complicated data structures with SOMs, including the assessment and monitoring of the topology preservation of the mapping. We distinguish a level of order in the SOM that is acceptable for cluster extraction. This can be achieved much earlier in the learning than finely tuned topography matching, but not as early as a sorted state. A sorted state, the mapping of known true classes in the SOM without scrambling does not guarantee successful detection of the same entities, because the prototypes may still not be molded sufficiently for a mU-matrix or other distance based similarity representations to align with the natural cluster boundaries. (Fig. 3, top center, is an example.) Our tool TopoView, presented in Sect. 3.1 can serve for the assessment of topology preservation on this level in relation to mU-matrix knowledge, as well as a verification tool for extracted clusters. The CONNvis SOM visualization, also a recent development, and our long time tool, the modified U-matrix, help achieve very detailed extraction of many relevant clusters, as shown in Sect. 4.2 and 4.3, representing dramatic improvement over some existing popular clustering capabilities such as ISODATA, for highly structured manifolds.

However, we point out that our tools could be further improved by combining the distance based knowledge of the mU-matrix and the topology based knowledge of the CONNvis. A natural extension will be to combine these two into one similarity measure, based on our experiences.

We do not discuss some aspects which could significantly contribute to SOM clustering but have not been much researched. Map magnification is one. This interesting subject is explored in [18] for highly structured data. Methods for verification of clusters (extracted by any algorithm) against the natural partitions in the data, are generally lacking for complicated data. Existing cluster

validity indices, which tend to favor particular types of data distributions (such as spherical clusters) fail to give accurate evaluation of the clustering quality for highly structured data. This is discussed in [60] and a new validity index, based on the same connectivity (CONN) matrix as used in CONNvis, is offered.

A valuable aspect of CONNvis SOM clustering is that it seems amenable to automation. Since the binning of connectivity stregths (line widths) in Fig. 9 is generated with thresholds derived automatically from the data characteristics as defined in [34], these thresholds can provide meaningful guidance for finding thinly textured parts of the manifold and cutting connections to achieve cluster separation. For this to work, however, the designation of coarse clusters by the human operator (as in Fig. 10) will need to be replaced by an automated consideration of the relationships between local and global connectedness at each prototype. While this is non-trivial we think it is doable and we are gathering insights from interactive CONNvis clustering for how to best implement this. Successful automation, with the same level of sophistication as shown here for interactive clustering, would significantly contribute to the solution of large problems such as on-board autonomous science, detection of small targets from unmanned vehicles in war zones, or precise mining of large security data bases.

Acknowledgments

We thank Drs. Juha Vesanto and Esa Alhoniemi for sharing their "Clown" data set and the SOM weights from their processing in [52]; Prof. John Kerekes, Rochester Intstitute of Technology, for providing the synthetic hyperspectral image used for this work, and Prof. Maj. Michael Mendenhall of the Air Force Institute of Technology, United States Air Force, for his help with preprocessing the same; Prof. Bea Csathó, University of Buffalo, for the Ocean City data and ground truth, as well as for ISODATA clusterings of the same. This work was partially supported by the Applied Information Systems Research Program (grant NNG05GA94G) of NASA's Science Mission Directorate.

References

1. Lee, J., Verleysen, M.: Nonlinear Dimensionality Reduction. Information Science and Statistics. Springer, New York (2007)
2. Gorban, A., Kégl, B., Wunsch, D., Zinovyev, A. (eds.): Principal Manifolds for data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering. Springer, New York (2008)
3. Cox, T.F., Cox, M.: Multidimensional Scaling. Chapman and Hall/CRC, Boca Raton (2001)
4. Tenenbaum, J.B., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
5. Roweis, S., Soul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
6. Donoho, D.L., Grimes, C.: Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Proc. National Academy of Sciences. 100, 5591–5596 (2003)

7. Pless, R.: Using Isomap to explore video sequences. In: Proc. International Conference on Computer Vision, pp. 1433–1440 (2003)
8. Yang, M.: Face Recognition Using Extended Isomap. In: Proc. International Conference on Image Processing ICIP 2002, vol. 2, pp. 117–120 (2002)
9. Polito, M., Perona, P.: Grouping and dimensionality reduction by locally linear embedding. In: Proc. Neural Information Processing Systems, NIPS (2001)
10. Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G., Koudas, N.: Non-linear dimensionality reduction techniques for classification and visualization. In: Proceedings of 8th SIGKDD, pp. 645–651 (2002)
11. Zhang, J., Li, S.Z., Wang, J.: Manifold learning and applications in recognition. In: Tan, Y.P., Kim Hui Yap, L.W. (eds.) Intelligent Multimedia Processing with Soft Computing. Springer, Heidelberg (2004)
12. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Heidelberg (1997)
13. Martinetz, T., Berkovich, S., Schulten, K.: Neural Gas network for vector quantization and its application to time-series prediction. IEEE Trans. on Neural Networks 4(4), 558–569 (1993)
14. Cottrell, M., Hammer, B., Hasenfuss, A., Villmann, T.: Batch and median neural gas. Neural Networks 19, 762–771 (2006)
15. Bishop, C.M., Svensen, M., Williams, C.K.I.: GTM: The Generative Topographic Mapping. Neural Computation 10(1), 215–234 (1998)
16. Aupetit, M.: 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, pp. 83–90. MIT Press, Cambridge (2006)
17. Bauer, H.U., Der, R., Herrmann, M.: Controlling the magnification factor of self–organizing feature maps. Neural Computation 8(4), 757–771 (1996)
18. Merényi, E., Jain, A., Villmann, T.: Explicit magnification control of self-organizing maps for "forbidden" data. IEEE Trans. on Neural Networks 18(3), 786–797 (2007)
19. Villmann, T., Claussen, J.: Magnification control in self-organizing maps and neural gas. Neural Computation 18, 446–469 (2006)
20. Hammer, B., Hasenfuss, A., Villmann, T.: Magnification control for batch neural gas. Neurocomputing 70, 1125–1234 (2007)
21. DeSieno, D.: Adding a conscience to competitive learning. In: Proc. IEEE Int'l Conference on Neural Networks (ICNN), New York, July 1988, vol. I, pp. I–117–124 (1988)
22. Cottrell, M., Fort, J., Pages, G.: Theoretical aspects of the SOM algorithm. Neurocomputing 21, 119–138 (1998)
23. Ritter, H., Schulten, K.: On the stationary state of Kohonen's self-organizing sensory mapping. Biol. Cyb. 54, 99–106 (1986)
24. Erwin, E., Obermayer, K., Schulten, K.: Self-organizing maps: ordering, convergence properties and energy functions. Biol. Cyb. 67, 47–55 (1992)
25. Hammer, B., Villmann, T.: Mathematical aspects of neural networks. In: Proc. Of European Symposium on Artificial Neural Networks (ESANN 2003), Brussels, Belgium. D facto publications (2003)
26. Martinetz, T., Schulten, K.: Topology representing networks. Neural Networks 7(3), 507–522 (1994)
27. Villmann, T., Der, R., Herrmann, M., Martinetz, T.: Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement. IEEE Transactions on Neural Networks 8(2), 256–266 (1997)
28. Bauer, H.U., Pawelzik, K.: Quantifying the neighborhood preservation of Self-Organizing Feature Maps. IEEE Trans. on Neural Networks 3, 570–579 (1992)

29. Kiviluoto, K.: Topology preservation in self-organizing maps. In: Proceedings IEEE International Conference on Neural Networks, Bruges, June 3–6, 1996, pp. 294–299 (1996)
30. Zhang, L., Merényi, E.: Weighted Differential Topographic Function: A Refinement of the Topographic Function. In: Proc. 14th European Symposium on Artificial Neural Networks (ESANN 2006), Brussels, Belgium, pp. 13–18. D facto publications (2006)
31. Csathó, B., Krabill, W., Lucas, J., Schenk, T.: A multisensor data set of an urban and coastal scene. In: Int'l Archives of Photogrammetry and Remote Sensing, vol. 32, pp. 26–31 (1998)
32. Bodt, E., Verleysen, M.C.: Statistical tools to assess the reliability of self-organizing maps. Neural Networks 15, 967–978 (2002)
33. Merényi, E., Tasdemir, K., Farrand, W.: Intelligent information extraction to aid science decision making in autonomous space exploration. In: Fink, W. (ed.) Proceedings of DSS 2008 SPIE Defense and Security Symposium, Space Exploration Technologies, Orlando, FL, Mach 17–18, 2008, vol. 6960, pp. 17–18. SPIE (2008) 69600M Invited
34. Tasdemir, K., Merényi, E.: Exploiting data topology in visualization and clustering of Self-Organizing Maps. IEEE Trans. on Neural Networks (2008) (in press)
35. Ultsch, A.: Self-organizing neural networks for visualization and classification. In: Opitz, O., Lausen, B. (eds.) Information and Classification — Concepts, Methods and Applications, pp. 307–313. Springer, Berlin (1993)
36. Kraaijveld, M., Mao, J., Jain, A.: A nonlinear projection method based on Kohonen's topology preserving maps. IEEE Trans. on Neural Networks 6(3), 548–559 (1995)
37. Merkl, D., Rauber, A.: Alternative ways for cluster visualization in Self-Organizing Maps. In: Proc. 1st Workshop on Self-Organizing Maps (WSOM 1997), Espoo, Finland, June 4-6, 1997, pp. 106–111 (1997)
38. Ultsch, A.: Maps for the visualization of high-dimensional data spaces. In: Proc. 4th Workshop on Self-Organizing Maps (WSOM 2003), Paris, France, vol. 3, pp. 225–230 (2003)
39. Cottrell, M., de Bodt, E.: A Kohonen map representation to avoid misleading interpretations. In: Proc. 4th European Symposium on Artificial Neural Networks (ESANN 1996), pp. 103–110. D-Facto, Bruges (1996)
40. Himberg, J.: A SOM based cluster visualization and its application for false colouring. In: Proc. IEEE-INNS-ENNS International Joint Conf. on Neural Networks, Como, Italy, vol. 3, pp. 587–592 (2000)
41. Kaski, S., Venna, J., Kohonen, T.: Coloring that reveals cluster structures in multivariate data. Australian Journal of Intelligent Information Processing Systems 6, 82–88 (2000)
42. Villmann, T., Merényi, E.: 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, pp. 121–145. Springer, Heidelberg (2001)
43. Vesanto, J.: SOM-Based Data Visualization Methods. Intelligent Data Analysis 3(2), 111–126 (1999)
44. Kaski, S., Kohonen, T., Venna, J.: Tips for SOM Processing and Colourcoding of Maps. In: Deboeck, G., Kohonen, T. (eds.) Visual Explorations in Finance Using Self-Organizing Maps, London (1998)

45. Pölzlbauer, G., Rauber, A., Dittenbach, M.: Advanced visualization techniques for self-organizing maps with graph-based methods. In: Jun, W., Xiaofeng, L., Zhang, Y. (eds.) Proc. Second Intl. Symp. on Neural Networks (ISSN 2005), Chongqing, China, pp. 75–80. Springer, Heidelberg (2005)
46. Aupetit, M., Catz, T.: High-dimensional labeled data analysis with topology representing graphs. Neurocomputing 63, 139–169 (2005)
47. Aupetit, M.: Visualizing the trustworthiness of a projection. In: Proc. 14th European Symposium on Artificial Neural Networks, ESANN 2006, Bruges, Belgium, April 26-28, 2006, pp. 271–276 (2006)
48. Howell, E.S., Merényi, E., Lebofsky, L.A.: Classification of asteroid spectra using a neural network. Jour. Geophys. Res. 99(E5), 10, 847–10, 865 (1994)
49. Merényi, E., Howell, E.S., et al.: Prediction of water in asteroids from spectral data shortward of 3 microns. ICARUS 129, 421–439 (1997)
50. Tasdemir, K., Merényi, E.: Considering topology in the clustering of self-organizing maps. In: Proc. 5th Workshop On Self-Organizing Maps (WSOM 2005), Paris, France, September 5–8, 2005, pp. 439–446 (2005)
51. Tasdemir, K., Merényi, E.: Data topology visualization for the Self-Organizing Map. In: Proc. 14th European Symposium on Artificial Neural Networks (ESANN 2006), Brussels, Belgium, April 26–28, 2006, pp. 125–130. D facto publications (2006)
52. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11(3), 586–600 (2000)
53. Merényi, E., Csató, B., Taşdemir, K.: Knowledge discovery in urban environments from fused multi-dimensional imagery. In: Gamba, P., Crawford, M. (eds.) Proc. IEEE GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (URBAN 2007), Paris, France, IEEE Catalog number 07EX1577, April 11-13, 2007, pp. 1–13 (2007)
54. Csathó, B., Schenk, T., Lee, D.C., Filin, S.: Inclusion of multispectral data into object recognition. Int'l Archives of Photogrammetry and Remote Sensing 32, 53–61 (1999)
55. Schott, J., Brown, S., Raqueño, R., Gross, H., Robinson, G.: An advanced synthetic image generation model and its application to multi/hyperspectral algorithm development. Canadian Journal of Remote Sensing 25(2) (June 1999)
56. Ientilucci, E., Brown, S.: Advances in wide-area hyperspectral image simulation. In: Proceedings of SPIE, May 5–8, 2003, vol. 5075, pp. 110–121 (2003)
57. Green, R.O.: Summaries of the 6th Annual JPL Airborne Geoscience Workshop, 1. In: AVIRIS Workshop, Pasadena, CA, March 4–6 (1996)
58. Green, R.O., Boardman, J.: Exploration of the relationship between information content and signal-to-noise ratio and spatial resolution. In: Proc. 9th AVIRIS Earth Science and Applications Workshop, Pasadena, CA, February 23–25 (2000)
59. Tou, J., Gonzalez, R.C.: Pattern Recognition Principles. Addison-Wesley Publishing Company, Reading (1974)
60. Tasdemir, K., Merényi, E.: A new cluster validity index for prototype based clustering algorithms based on inter- and intra-cluster density. In: Proc. Int'l Joint Conf. on Neural Networks (IJCNN 2007), Orlando, FL, August 12–17, 2007, pp. 2205–2211 (2007)