ARTICLE IN PRESS

PERGAMON

Neural Networks xx (0000) 1–15

Neural Networks

www.elsevier.com/locate/neunet

Neural maps in remote sensing image analysis

Thomas Villmann[a,*], Erzsébet Merényi[b], Barbara Hammer[c]

[a]Klinik für Psychotherapie, Universität Leipzig, Karl-Tauchnitz-Str. 25, 04107 Leipzig, Germany
[b]Department of Electrical and Computer Engineering, Rice University, 6100 Main Street, MS 380 Houston, TX, USA
[c]Department of Mathematics/Computer Science, University of Osnabrück, Albrechtstraße 28, 49069 Osnabrück, Germany

Abstract

We study the application of self-organizing maps (SOMs) for the analyses of remote sensing spectral images. Advanced airborne and satellite-based imaging spectrometers produce very high-dimensional spectral signatures that provide key information to many scientific investigations about the surface and atmosphere of Earth and other planets. These new, sophisticated data demand new and advanced approaches to cluster detection, visualization, and supervised classification. In this article we concentrate on the issue of faithful topological mapping in order to avoid false interpretations of cluster maps created by an SOM. We describe several new extensions of the standard SOM, developed in the past few years: the growing SOM, magnification control, and generalized relevance learning vector quantization, and demonstrate their effect on both low-dimensional traditional multi-spectral imagery and ~200-dimensional hyperspectral imagery.

Keywords: Self-organizing map; Remote sensing; Generalized relevance learning vector quantization

1. Introduction

In geophysics, geology, astronomy, as well as in many environmental applications air- and satellite-borne remote sensing spectral imaging has become one of the most advanced tools for collecting vital information about the surface of Earth and other planets. Spectral images consist of an array of multi-dimensional vectors each of which is assigned to a particular spatial area, reflecting the response of a spectral sensor for that area at various wavelengths (Fig. 1). Classification of intricate, high-dimensional spectral signatures has turned out to be far from trivial. Discrimination among many surface cover classes, discovery of spatially small, interesting spectral species proved to be an insurmountable challenge to many traditional clustering and classification methods. By customary measures (such as, for example, principal component analysis (PCA) the intrinsic spectral dimensionality of hyperspectral images appears to be surprisingly low, 5–10 at most. Yet dimensionality reduction to such low numbers has not been successful in terms of preservation of important class distinctions. The spectral bands, many of which are highly correlated, may lie on a low-dimensional but non-linear manifold, which is a scenario that eludes many classical approaches. In addition, these data comprise enormous volumes and are frequently noisy. This motivates research into advanced and novel approaches for remote sensing image analysis and, in particular, neural networks (Merényi, 1999). Generally, artificial neural networks attempt to replicate the *computational power* of biological neural networks, with the following important (incomplete list of) features:

- *adaptivity*—the ability to change the internal representation (connection weights, network structure) if new data or information are available;
- *robustness*—handling of missing, noisy ore otherwise confused data;
- *power/speed*—handling of large data volumes in acceptable time due to inherent parallelism;
- *non-linearity*—the ability to represent non-linear functions or mappings.

Exactly these properties make the application of neural networks interesting in remote sensing image analysis. In the present contribution, we will concentrate on a special neural network type, neural maps. Neural maps constitute an important neural network paradigm. In brains, neural maps occur in all sensory modalities as well as in motor areas. In technical contexts, neural maps are utilized in the fashion of neighborhood preserving vector quantizers. In both cases

* Corresponding author. Tel.: +49-341-9718868; fax: +49-341-2131257.
E-mail address: villmann@informatik.uni-leipzig.de (T. Villmann).

0893-6080/03/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S0893-6080(03)00021-2

NN 1720—1/3/2003—13:50—SUREKHA—64643— MODEL 5

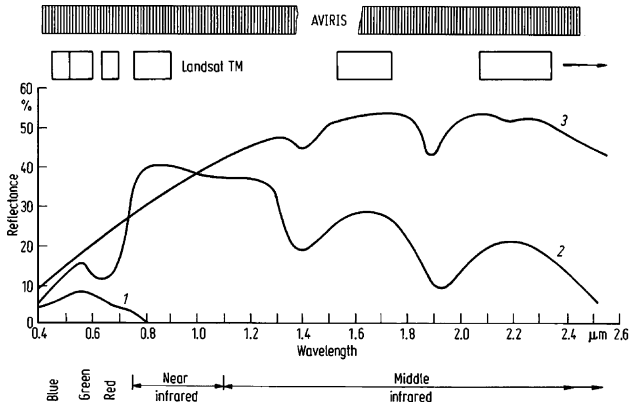

Fig. 1. Comparison of the spectral resolution of the AVIRIS hyperspectral imager and the LANDSAT TM imager in the visible and near-infrared spectral range. AVIRIS represents a quasi-continuous spectral sampling whereas LANDSAT TM has coarse spectral resolution with large gaps between the band passes. Additionally, spectral reflectance charateristics of common earth surface materials is shown: 1, water; 2, vegetation; 3, soil (from Richards and Jia (1999)).

these networks project data from some *possibly high-dimensional* input space onto a position in some output space. In the area of spectral image analysis we apply neural maps for clustering, data dimensionality reduction, learning of relevant data dimensions and visualization. For this purpose we highlight several useful extensions of the basic neural map approaches.

This paper is structured as follows. In Section 2 we briefly introduce the problems in remote sensing spectral image analysis. In Section 3 we give the basics of neural maps and vector quantization together with a short review of new extensions: structure adaptation and magnification control (Section 3.2), and learning of proper metrics or relevance learning (Section 3.3). In Section 4 we present applications of the reviewed methods to multi-spectral LANDSAT TM imagery as well as very high-dimensional hyperspectral AVIRIS imagery.

2. Remote sensing spectral imaging

As mentioned earlier air- and satellite-borne remote sensing spectral images consist of an array of multi-dimensional vectors assigned to particular spatial regions (pixel locations) reflecting the response of a spectral sensor at various wavelengths. These vectors are called spectra. A spectrum is a characteristic pattern that provides a clue to the surface material within the respective surface element. The utilization of these spectra includes areas such as mineral exploration, land use, forestry, ecosystem management, assessment of natural hazards, water resources, environmental contamination, biomass and productivity; and many other activities of economic significance, as well as prime scientific pursuits such as looking for possible sources of past or present life on other planets. The number of applications has dramatically increased in the past 10 years with the advent of imaging spectrometers such as AVIRIS of NASA/JPL, that greatly surpass the traditional multi-spectral sensors (e.g. LANDSAT thematic mapper (TM).

Imaging spectrometers can resolve the known, unique, discriminating spectral features of minerals, soils, rocks, and vegetation. While a multi-spectral sensor samples a given wavelength window (typically the $0.4-2.5\ \mu m$ range in the case of visible and near-infrared imaging) with several broad bandpasses, leaving large gaps between the bands, imaging spectrometers sample a spectral window contiguously with very narrow, $10-20$ nm badpasses (Fig. 1) (Richards & Jia, 1999). Hyperspectral technology is in great demand because direct identification of surface compounds is possible

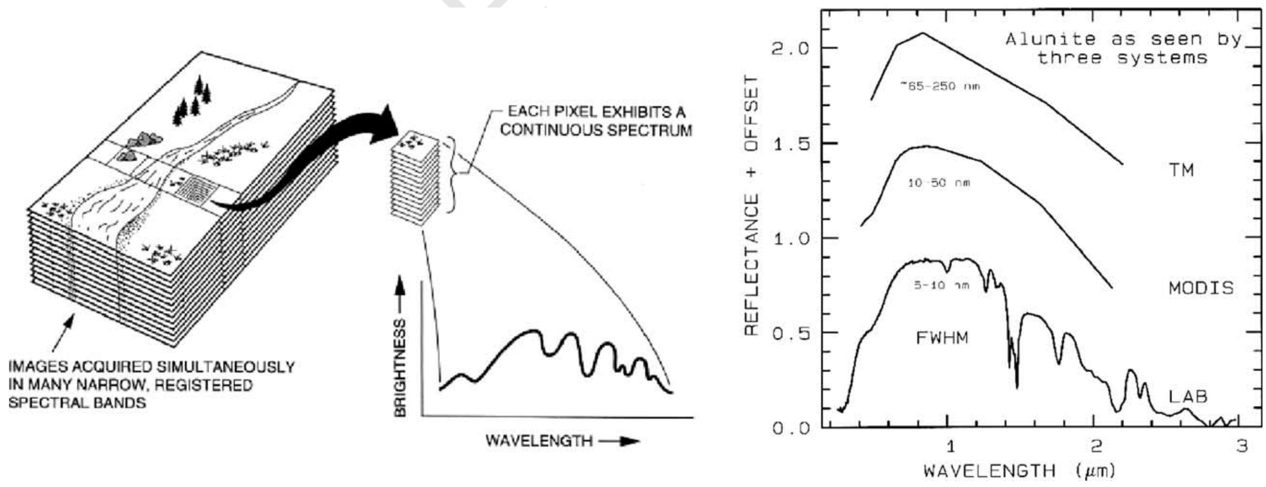

Fig. 2. Left: The concept of hyperspectral imaging. Figure from Campbell (1996). Right: The spectral signature of the mineral alunite as seen through the six broad bands of Landsat TM, as seen by the moderate spectral resolution sensor MODIS (20 bands in this region), and as measured in laboratory. Figure from Clark (1999). Hyperspectral sensors such as AVIRIS of NASA/JPL (Green, 1996) produce spectral details comparable to laboratory measurements.

without prior field work, for materials with known spectral signatures. Depending on the wavelength resolution and the width of the wavelength window used by a particular sensor, the dimensionality of the spectra can be as low as 5–6 (such as in LANDSAT TM), or as high as several hundred for hyperspectral imagers (Green, 1996).

Spectral images can formally be described as a matrix $\mathbf{S} = \mathbf{v}^{(x,y)}$, where $\mathbf{v}^{(x,y)} \in \mathbb{R}^{D_{\mathscr{V}}}$ is the vector of spectral information associated with pixel location (x, y). The elements $v_i^{(x,y)}$, $i = 1 \ldots D_{\mathscr{V}}$ of spectrum $\mathbf{v}^{(x,y)}$ reflect the responses of a spectral sensor at a suite of wavelengths (Fig. 2; Campbell, 1996; Clark, 1999). The spectrum is a characteristic fingerprint pattern that identifies the surface material within the area defined by pixel (x, y). The individual two-dimensional image $\mathbf{S}_i = v_i^{(x,y)}$ at wavelength i is called the ith image band. The data space $\mathscr{V}$ spanned by visible–near infrared reflectance spectra is $[0 - \text{noise}, U + \text{noise}]^{D_{\mathscr{V}}} \subseteq \mathbb{R}^{D_{\mathscr{V}}}$ where $U > 0$ represents an upper limit of the measured scaled reflectivity and noise is the maximum value of noise across all spectral channels and image pixels. The data density $\mathscr{P}(\mathscr{V})$ may vary strongly within this space. Sections of the data space can be very densely populated while other parts may be extremely sparse, depending on the materials in the scene and on the spectral bandpasses of the sensor. According to this model traditional multi-spectral imagery has a low $D_{\mathscr{V}}$ value while $D_{\mathscr{V}}$ can be several hundred for hyperspectral images. The latter case is of particular interest because the great spectral detail, complexity, and very large data volume pose new challenges in clustering, cluster visualization, and classification of images with such high spectral dimensionality (Merényi, 1998).

In addition to dimensionality and volume, other factors, specific to remote sensing, can make the analyses of hyperspectral images even harder. For example, given the richness of data, the goal is to separate many cover classes, however, surface materials that are significantly different for an application may be distinguished by very subtle differences in their spectral patterns. The pixels can be mixed, which means that several different materials may contribute to the spectral signature associated with one pixel. Training data may be scarce for some classes, and classes may be represented very unevenly. All these difficulties motivate research into advanced and novel approaches (Merényi, 1999).

Noise is far less problematic than the intricacy of the spectral patterns, because of the high signal-to-noise ratios (500–1500) that present-day hyperspectral imagers provide. For this discussion, we will omit noise issues, and additional effects such as atmospheric distortions, illumination geometry and albedo variations in the scene, because these can be addressed through well-established procedures prior to clustering or classification.

3. Neural maps and vector quantization

3.1. Basic concepts and notations

Neural maps as tools for (unsupervised) topographic vector quantization algorithms map data vectors $\mathbf{v}$ sampled from a data manifold $\mathscr{V} \subseteq \mathbb{R}^{D_{\mathscr{V}}}$ onto a set $\mathscr{A}$ of neurons $\mathbf{r}$

$$\Psi_{\mathscr{V} \to \mathscr{A}} : \mathscr{V} \to \mathscr{A}. \tag{3.1}$$

Associated with each neuron is a pointer (weight vector) $\mathbf{w}_\mathbf{r} \in \mathbb{R}^{D_{\mathscr{V}}}$ specifying the configuration of the neural map $\mathbf{W} = \{\mathbf{w}_\mathbf{r}\}_{\mathbf{r} \in \mathscr{A}}$. The task of vector quantization is realized by the map Ψ as a winner-take-all rule, i.e. a stimulus vector $\mathbf{v} \in \mathscr{V}$ is mapped onto that neuron $\mathbf{s} \in \mathscr{A}$ the pointer $\mathbf{w}_s$ of which is closest to the presented stimulus vector $\mathbf{v}$,

$$\Psi_{\mathscr{V} \to \mathscr{A}} : \mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{r} \in \mathscr{A}} d(\mathbf{v}, \mathbf{w}_\mathbf{r}). \tag{3.2}$$

where $d(\mathbf{v}, \mathbf{w}_\mathbf{r})$ is the Euclidean distance. The neuron $\mathbf{s}$ is referred to as the *winner* or *best-matching unit*. The reverse mapping is defined as $\Psi_{\mathscr{A} \to \mathscr{V}} : \mathbf{r} \mapsto \mathbf{w}_\mathbf{r}$. These two mappings together determine the neural map

$$\mathscr{M} = (\Psi_{\mathscr{V} \to \mathscr{A}}, \Psi_{\mathscr{A} \to \mathscr{V}}), \tag{3.3}$$

realized by the network. The subset of the input space

$$\Omega_\mathbf{r} = \{\mathbf{v} \in \mathscr{V} : \mathbf{r} = \Psi_{\mathscr{V} \to \mathscr{A}}(\mathbf{v})\}, \tag{3.4}$$

which is mapped to a particular neuron $\mathbf{r}$ according to Eq. (3.2), forms the (masked) Ω receptive field of that neuron. If the intersection of two masked receptive fields $\Omega_\mathbf{r}$, $\Omega_{\mathbf{r}'}$ is non-vanishing we call $\Omega_\mathbf{r}$ and $\Omega_{\mathbf{r}'}$ neighbored. The neighborhood relations form a corresponding graph structure $\mathscr{G}_{\mathscr{V}}$ in $\mathscr{A}$: two neurons are connected in $\mathscr{G}_{\mathscr{V}}$ if and only if their masked receptive fields are neighbored. The graph $\mathscr{G}_{\mathscr{V}}$ is called the induced Delaunay-graph (See, for example, Martinetz and Schulten (1994) for detailed definitions). Due to the bijective relation between neurons and weight vectors, $\mathscr{G}_{\mathscr{V}}$ also represents the Delaunay graph of the weights (Martinetz & Schulten, 1994).

In the most widely used neural map algorithm, the self-organizing map (SOM) (Kohonen, 1995), the neurons are arranged a priori on a fixed grid $\mathscr{A}$.[1] Other algorithms such as the topology representing network (TRN), which is based on the neural gas algorithm (NG) (Martinetz, Berkovich, & Schulten, 1993), do not specify the topological relations among neurons, but construct a neighborhood structure in the output space during learning (Martinetz & Schulten, 1994).

In any of these algorithms the pointers are adapted such that the input data are matched appropriately. For this purpose a random sequence of data points $\mathbf{v} \in \mathscr{V}$ from the stimuli distribution $\mathscr{P}(\mathbf{v})$ is presented to the

[1] Usually the grid $\mathscr{A}$ is chosen as a $d_{\mathscr{A}}$-dimensional hypercube and then one has $i = (i_1, \ldots, i_{d_{\mathscr{A}}})$. Yet, other ordered arrangements are also admissible.

map, the winning neuron $\mathbf{s}$ is determined according to Eq. (3.2), and the pointer $\mathbf{w_s}$ is shifted toward $\mathbf{v}$. Interactions among the neurons are included via a neighborhood function determining to what extent the units that are close to the winner in the output space participate in the learning step

$$\Delta\mathbf{w_r} = \epsilon h_\lambda(\mathbf{r},\mathbf{s},\mathbf{v})(\mathbf{v} - \mathbf{w_r}). \tag{3.5}$$

$h_\lambda(\mathbf{r},\mathbf{s},\mathbf{v})$ is usually chosen to be of Gaussian shape

$$h_\lambda(\mathbf{r},\mathbf{s},\mathbf{v}) = \exp\left(-\frac{(d_{\mathscr{A}}(\mathbf{r},\mathbf{s}(\mathbf{v})))^2}{\lambda}\right), \tag{3.6}$$

where $d_{\mathscr{A}}(\mathbf{r},\mathbf{s}(\mathbf{v}))$ is a distance measure on the set $\mathscr{A}$. We emphasize that $h_\lambda(\mathbf{r},\mathbf{s},\mathbf{v})$ implicitly depends on the whole set $\mathbf{W}$ of weight vectors via the winner determination of $\mathbf{s}$ according to Eq. (3.2). In SOMs it is evaluated in the output space $\mathscr{A}: d_{\mathscr{A}}^{\mathrm{SOM}}(\mathbf{r},\mathbf{s}(\mathbf{v})) = \|\mathbf{r} - \mathbf{s}(\mathbf{v})\|_{\mathscr{A}}$ whereas for the NG we have $d_{\mathscr{A}}^{\mathrm{NG}}(\mathbf{r},\mathbf{s}(\mathbf{v})) = k_\mathbf{r}(\mathbf{v},\mathbf{W})$. $k_\mathbf{r}(\mathbf{v},\mathbf{W})$ gives the number of pointers $\mathbf{w}_{\mathbf{r}'}$ for which the relation $d(\mathbf{v},\mathbf{w}_{\mathbf{r}'}) \le d(\mathbf{v},\mathbf{w_r})$ is valid (Martinetz et al., 1993), i.e. $d_{\mathscr{A}}^{\mathrm{NG}}$ is evaluated in the input space (Here, $d(,)$ is the Euclidean distance.). In the SOM usually $\lambda = 2\sigma^2$ is used.

The particular definitions of the distance measures in the two models cause further differences. In Martinetz et al. (1993) the convergence rate of the neural gas network was shown to be faster than that of the SOM. Furthermore, it was proven that the adaptation rule for the weight vectors follows, on average, a potential dynamic. In contrast, a global energy function does not exist in the SOM algorithm (Erwin, Obermayer, & Schulten, 1992).

One important extension of the basic concepts of neural maps concerns the so-called *magnification*. The standard SOM distributes the pointers according to the input distribution

$$\mathscr{P}(\mathbf{W}) \sim \mathscr{P}(\mathscr{V})^\alpha, \tag{3.7}$$

with the magnification factor $\alpha_{\mathrm{SOM}} = \frac{2}{3}$ (Ritter & Schulten, 1986; Kohonen, 1999).[2,3] For the NG one finds, by analytical considerations, that for small λ the magnification factor $\alpha_{\mathrm{NG}} = d/(d+2)$ which only depends on the dimensionality of the input space embedded in $\mathbb{R}^{D_{\mathscr{V}}}$, i.e. the result is valid for all dimensions (Martinetz et al., 1993).

Topology preservation or topographic mapping in neural maps is defined as the preservation of the continuity of the mapping from the input space onto the output space, more precisely it is equivalent to the *continuity* of $\mathscr{M}$ between the *topological spaces* with properly chosen metric in both $\mathscr{A}$ and $\mathscr{V}$. For lack of space we refer to Villmann, Der, and Herrmann (1997) for detailed considerations. For the TRN the metric in $\mathscr{A}$ is defined according to the given data structure whereas in case of the SOM violations of topology preservation may occur if the structure of the output space does not match the topological structure of $\mathscr{V}$. In SOMs the topology preserving property can be used for immediate evaluations of the resulting map, for instance by interpretation as a color space, as demonstrated in Section 4.2.2. Topology preservation also allows the applications of interpolating schemes such as the parametrized SOM (PSOM) (Ritter, 1993) or interpolating SOM (I-SOM) (Goppert & Rosenstiel, 1997). A higher degree of topology preservation, in general, improves the accuracy of the map (Bauer & Pawelzik, 1992).

As pointed out in Villmann et al. (1997) violations of topographic mapping in SOMs can result in false interpretations. Several approaches were developed to judge the degree of topology preservation for a given map. Here we briefly describe a variant $\tilde{P}$ of the well known topographic product P (Bauer & Pawelzik, 1992). Instead of the Euclidean distances between the weight vectors, $\tilde{P}$ uses the respective distances $d^{\mathscr{G}_{\mathscr{V}}}(\mathbf{w_r}, \mathbf{w}_{\mathbf{r}'})$ of minimal path lengths in the induced Delaunay-graph $\mathscr{G}_{\mathscr{V}}$ of the $\mathbf{w_r}$. During the computation of $\tilde{P}$ for each node $\mathbf{r}$ the sequences $\mathbf{n}_j^{\mathscr{A}}(\mathbf{r})$ of jth neighbors of $\mathbf{r}$ in $\mathscr{A}$, and $\mathbf{n}_j^{\mathscr{V}}(\mathbf{r})$ describing the jth neighbor of $\mathbf{w_r}$ in $\mathscr{V}$, have to be determined. These sequences and further averaging over neighborhood orders j and nodes $\mathbf{r}$ finally lead to

$$\tilde{P} = \frac{1}{N(N-1)}\sum_{\mathbf{r}}\sum_{j=1}^{N-1}\frac{1}{2j}\log(\Theta), \tag{3.8}$$

with

$$\Theta = \Pi_{l=1}^{j}\frac{d^{\mathscr{G}_{\mathscr{V}}}(\mathbf{w_r},\mathbf{w}_{\mathbf{n}_l^{\mathscr{A}}(\mathbf{r})})}{d^{\mathscr{G}_{\mathscr{V}}}(\mathbf{w_r},\mathbf{w}_{\mathbf{n}_l^{\mathscr{V}}(\mathbf{r})})}\,\frac{d_{\mathscr{A}}(\mathbf{r},\mathbf{n}_l^{\mathscr{A}}(\mathbf{r}))}{d_{\mathscr{A}}(\mathbf{r},\mathbf{n}_l^{\mathscr{V}}(\mathbf{r}))}, \tag{3.9}$$

and $d_{\mathscr{A}}(\mathbf{r},\mathbf{r}')$ is the distance between the neuron positions $\mathbf{r}$ and $\mathbf{r}'$ in the lattice $\mathscr{A}$. $\tilde{P}$ can take on positive or negative values with similar meaning as for the original topographic product P: if $\tilde{P} < 0$ holds the output space is too low-dimensional, and for $\tilde{P} > 0$ the output space is too high-dimensional. In both cases neighborhood relations are violated. Only for $\tilde{P} \approx 0$ does the output space approximately match the topology of the input data. The present variant $\tilde{P}$ overcomes the problem of strongly curved maps which may be judged neighborhood violating by the original P even though the shape of the map might be perfectly justified (Villmann et al., 1997).

---

[2] This result is valid for the one-dimensional case and higher-dimensional ones which separate.

[3] The notation 'magnification factor' is, strictly speaking, not correct. Some times it is called (mathematically exact) *magnification exponent*. However, the notation 'magnification factor' is commonly used (van Hulle, 2000). Therefore we use this notation here.

3.2. Magnification control and structure adaptation

3.2.1. Magnification control

The first approach to influence the magnification of a vector quantizer, proposed in DeSieno (1988), is called the *mechanism of conscience* yielding approximately the same winning probability for each neuron. Hence, the approach yields density matching between input and output space which corresponds to a magnification factor of 1. Bauer, Der, and Herrmann (1996) suggested a more general scheme for the SOM to achieve an arbitrary magnification. They introduced a local learning parameter $\epsilon_{\mathbf{s}}$ for each neuron $\mathbf{r}$ with $\langle\epsilon_{\mathbf{s}}\rangle \propto \mathscr{P}(\mathbf{w_r})^m$ in Eq. (3.5), where m is an additional control parameter. Eq. (3.5) now reads as

$$\Delta\mathbf{w_r} = \epsilon_{\mathbf{s}} h_\lambda(\mathbf{r},\mathbf{s},\mathbf{v})(\mathbf{v}-\mathbf{w_r}). \tag{3.10}$$

Note that the learning factor $\epsilon_{\mathbf{s}}$ of the winning neuron $\mathbf{s}$ is applied to all updates. This local learning leads to a similar relation as in Eq. (3.7)

$$\mathscr{P}(\mathbf{W}) \sim \mathscr{P}(\mathscr{V})^{\alpha'}, \tag{3.11}$$

with $\alpha' = \alpha(m+1)$ and allows a magnification control through the choice of m. In particular, one can achieve a resolution of $\alpha' = 1$, which maximizes mutual information (Linsker, 1989; Villmann & Herrmann, 1998). Application of this approach to the derivation of the dynamic of TRN yields exactly the same result (Villmann, 2000).

3.2.2. Structure adaptation

For both neural map types, the SOM and the TRN, growing variants for structure adaptation are known.

In the growing SOM (GSOM) approach the output $\mathscr{A}$ is a hypercube in which both the dimension and the respective length ratios are adapted during the learning procedure in addition to the weight vector learning, i.e. the overall dimensionality and the dimensions along the individual directions change in the hypercube. The GSOM starts from an initial 2-neuron chain, learns according the regular SOM-algorithm, adds neurons to the output space according to a certain criterion to be described later, learns again, adds again, etc. until a prespecified maximum number $N_{\max}$ of neurons is distributed. During this procedure, the output space topology remains of the form $n_1 \times n_2 \times \cdots$, with $n_j = 1$ for $j > D_{\mathscr{A}}$, where $D_{\mathscr{A}}$ is the current dimensionality of $\mathscr{A}$.[4] From there it can grow either by adding nodes in one of the directions which are already spanned by the output space or by initialising a new dimension. This decision is made on the basis of the receptive fields $\Omega_{\mathbf{r}}$. When reconstructing $\mathbf{v} \in \mathscr{V}$ from neuron $\mathbf{r}$, an error $\boldsymbol{\theta} = \mathbf{v} - \mathbf{w_r}$ remains decomposed along the different directions, which results from projecting back the output grid $\mathscr{A}$ into the input space $\mathscr{V}$

$$\boldsymbol{\theta} = \mathbf{v} - \mathbf{w_r} = \sum_{i=1}^{D_{\mathscr{A}}} a_i(\mathbf{v}) \frac{\mathbf{w}_{\mathbf{r}+\mathbf{e}_i} - \mathbf{w}_{\mathbf{r}-\mathbf{e}_i}}{\|\mathbf{w}_{\mathbf{r}+\mathbf{e}_i} - \mathbf{w}_{\mathbf{r}-\mathbf{e}_i}\|} + \mathbf{v}'. \tag{3.12}$$

Here, $\mathbf{e}_i$ denotes the unit vector in direction i of $\mathscr{A}$, $\mathbf{w}_{\mathbf{r}+\mathbf{e}_i}$ and $\mathbf{w}_{\mathbf{r}-\mathbf{e}_i}$ are the weight vectors of the neighbors of the neuron $\mathbf{r}$ in the ith direction of the lattice $\mathscr{A}$.[5] Considering a receptive field $\Omega_{\mathbf{r}}$ and determining the first principle component $\boldsymbol{\omega}_{\mathrm{PCA}}$ of $\Omega_{\mathbf{r}}$ allows a further decomposition of $\mathbf{v}'$. Projection of $\mathbf{v}'$ onto the direction of $\boldsymbol{\omega}_{\mathrm{PCA}}$ then yields $a_{D_{\mathscr{A}}+1}(\mathbf{v})$,

$$\mathbf{v}' = a_{D_{\mathscr{A}}+1}(\mathbf{v}) \frac{\boldsymbol{\omega}_{\mathrm{PCA}}}{\|\boldsymbol{\omega}_{\mathrm{PCA}}\|} + \mathbf{v}''. \tag{3.13}$$

The *criterion for the growing* now is to add nodes in that direction which has, on average, the largest of the expected (normalized) error amplitudes $\tilde{a}_i$

$$\tilde{a}_i = \sqrt{\frac{n_i}{n_i+1}} \sum_{\mathbf{v}\in\Omega_{\mathbf{r}}} \frac{|a_i(\mathbf{v})|}{\sqrt{\sum_{j=1}^{D_{\mathscr{A}}+1} a_j^2(\mathbf{v})}}, \tag{3.14}$$

$$i = 1, \ldots, D_{\mathscr{A}} + 1.$$

After each growth step, a new learning phase has to take place, in order to readjust the map. For a detailed study of the algorithm we refer to Bauer and Villmann (1997).

The growing TRN adapts the number of neurons in TRN whereby the structure between them is re-arranged according to the TRN definition (Fritzke, 1995). Hence, it is capable of representing the data space in a topologically faithful manner with increasing accuracy and it also realizes a structure adaptation.

3.3. Relevance learning

As mentioned earlier neural maps are unsupervised learning algorithms. In case of labeled data one can apply supervised classification schemes for vector quantization (VQ). Assuming that the data are labeled, an automatic clustering can be learned via attaching maps to the SOM or adding a supervised component to VQ to achieve learning vector quantization (LVQ) (Kohonen, Lappalainen, & Saljärvi, 1997; Meyering & Ritter, 1992). Various modifications of LVQ exist which ensure faster convergence, a better adaptation of the receptive fields to optimum Bayesian decision, or an adaptation for complex data

[4] Hence, the initial configuration is $2 \times 1 \times 1 \times \cdots, D_{\mathscr{A}} = 1$.

[5] At the border of the output space grid, where not two, but just one neighboring neuron is available, we use

$$\frac{\mathbf{w_r} - \mathbf{w}_{\mathbf{r}-\mathbf{e}_i}}{\|\mathbf{w_r} - \mathbf{w}_{\mathbf{r}-\mathbf{e}_i}\|}$$

or

$$\frac{\mathbf{w}_{\mathbf{r}+\mathbf{e}_i} - \mathbf{w_r}}{\|\mathbf{w}_{\mathbf{r}-\mathbf{e}_i} - \mathbf{w_r}\|}$$

to compute the backprojection of the output space direction $\mathbf{e}_i$ into the input space.

structures, to mention just a few (Kohonen et al., 1997; Sato & Yamada, 1995; Somervuo & Kohonen, 1999).

A common feature of unsupervised algorithms and LVQ is the information provided by the distance structure in $\mathscr{V}$, which is determined by the chosen metric. Learning heavily relies on this feature and therefore crucially depends on how appropriate the chosen metric is for the respective learning task. Several methods have been proposed which adapt the metric during training (Kaski, 1998; Kaski & Sinkkonen, 1999; Pregenzer, Pfurtscheller, & Flotzinger, 1996).

Here we focus on metric adaptation for LVQ since the LVQ combines the elegance of simple and intuitive updates in unsupervised algorithms with the accuracy of supervised methods. Let $c_{\mathbf{v}} \in \mathscr{L}$ be the label of input $\mathbf{v}$, $\mathscr{L}$ a set of labels (classes) with $\#\mathscr{L} = N_{\mathscr{L}}$. LVQ uses a fixed number of codebook vectors (weight vectors) for each class. Let $\mathbf{W} = \{\mathbf{w_r}\}$ be the set of all codebook vectors and $c_{\mathbf{r}}$ be the class label of $\mathbf{w_r}$. Furthermore, let $\mathbf{W}_c = \{\mathbf{w_r} | c_{\mathbf{r}} = c\}$ be the subset of weights assigned to class $c \in \mathscr{L}$.

The training algorithm adapts the codebook vectors such that for each class $c \in \mathscr{L}$, the corresponding codebook vectors $\mathbf{W}_c$ represent the class as accurately as possible. This means that the set of points in any given class $\mathscr{V}_c = \{\mathbf{v} \in \mathscr{V} | c_{\mathbf{v}} = c\}$, and the union $\mathscr{U}_c = \bigcup_{\mathbf{r}|_{\mathbf{w_r} \in \mathbf{W}_c}} \Omega_{\mathbf{r}}$ of receptive fields of the corresponding codebook vectors should differ as little as possible. For a given data point $\mathbf{v}$ with class label $c_{\mathbf{v}}$ let $\mu(\mathbf{v})$ denote some function which is negative if $\mathbf{v}$ is classified correctly, i.e. it belongs to a receptive field $\Omega_{\mathbf{r}}$ with $c_{\mathbf{r}} = c_{\mathbf{v}}$, and which is positive if $\mathbf{v}$ is classified incorrectly, i.e. it belongs to a receptive field $\Omega_{\mathbf{r}}$ with $c_{\mathbf{r}} \neq c_{\mathbf{v}}$. Let $f : \mathbb{R} \to \mathbb{R}$ be some monotonically increasing function. The general scheme of GLVQ aims to minimize the cost function

$$\mathrm{Cost} = \sum_{\mathbf{v}} f(\mu(\mathbf{v})), \tag{3.15}$$

via stochastic gradient descent. Different choices of f and $\mu(\mathbf{v})$ yield LVQ or LVQ2.1 as introduced in Kohonen et al. (1997), respectively, as shown in Sato and Yamada (1995). Denote by d the squared Euclidean distance of $\mathbf{v}$ to the nearest codebook vector. The choice of f as the sigmoidal function and

$$\mu(\mathbf{v}) = \frac{d_{\mathbf{r}_+} - d_{\mathbf{r}_-}}{d_{\mathbf{r}_+} + d_{\mathbf{r}_-}}, \tag{3.16}$$

where $d_{\mathbf{r}_+}$ is the squared Euclidean distance of the input vector $\mathbf{v}$ to the nearest codebook vector labeled with $c_{\mathbf{r}_+} = c_{\mathbf{v}}$, say $\mathbf{w}_{\mathbf{r}_+}$, and $d_{\mathbf{r}_-}$ is the squared Euclidean distance to the best matching codebook vector but labeled with $c_{\mathbf{r}_-} \neq c_{\mathbf{r}_+}$, say $\mathbf{w}_{\mathbf{r}_-}$, yields a powerful and noise tolerant behavior since it combines adaptation near the optimum Bayesian borders similarly as in LVQ2.1, with prohibiting the possible divergence of LVQ2.1 (as reported in Sato and Yamada (1995)). We refer to this modification as the GLVQ. The respective learning dynamic is obtained by differentiation of the cost function (3.15) (Sato & Yamada, 1995)

$$\Delta \mathbf{w}_{\mathbf{r}_+} = \epsilon \frac{4 f'|_{\mu(\mathbf{v})} \cdot d_{\mathbf{r}_-}}{(d_{\mathbf{r}_+} + d_{\mathbf{r}_-})^2} (\mathbf{v} - \mathbf{w}_{\mathbf{r}_+}) \quad \text{and}$$
$$\Delta \mathbf{w}_{\mathbf{r}_-} = -\epsilon \frac{4 f'|_{\mu(\mathbf{v})} \cdot d_{\mathbf{r}_+}}{(d_{\mathbf{r}_+} + d_{\mathbf{r}_-})^2} (\mathbf{v} - \mathbf{w}_{\mathbf{r}_-}). \tag{3.17}$$

To result a metric adaptation we now introduce input weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{D_{\mathscr{V}}})$, $\lambda_i \geq 0$, $\|\boldsymbol{\lambda}\| = 1$ in order to allow a different scaling of the input dimensions: substituting the squared Euclidean metric by its scaled variant

$$d^{\boldsymbol{\lambda}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D_{\mathscr{V}}} \lambda_i (x_i - y_i)^2, \tag{3.18}$$

the receptive field of codebook vector $\mathbf{w_r}$ becomes

$$\Omega_{\mathbf{r}}^{\boldsymbol{\lambda}} = \{\mathbf{v} \in \mathscr{V} : \mathbf{r} = \Psi^{\boldsymbol{\lambda}}_{\mathscr{V} \to \mathscr{A}}(\mathbf{v})\}, \tag{3.19}$$

with

$$\Psi^{\boldsymbol{\lambda}}_{\mathscr{V} \to \mathscr{A}} : \mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{r} \in \mathscr{A}} d^{\boldsymbol{\lambda}}(\mathbf{v}, \mathbf{w_r}), \tag{3.20}$$

as mapping rule. Replacing $\Omega_{\mathbf{r}}$ by $\Omega_{\mathbf{r}}^{\boldsymbol{\lambda}}$ and d by $d^{\boldsymbol{\lambda}}$ in the cost function 'Cost' in Eq. (3.15) yields a variable weighting of the input dimensions and hence an *adaptive metric*. Now Eq. (3.17) reads as

$$\Delta \mathbf{w}_{\mathbf{r}_+} = \epsilon \frac{4 \cdot f'|_{\mu(\mathbf{v})} \cdot d^{\boldsymbol{\lambda}}_{\mathbf{r}_-}}{(d^{\boldsymbol{\lambda}}_{\mathbf{r}_+} + d^{\boldsymbol{\lambda}}_{\mathbf{r}_-})^2} \boldsymbol{\Lambda} (\mathbf{v} - \mathbf{w}_{\mathbf{r}_+}) \quad \text{and}$$
$$\Delta \mathbf{w}_{\mathbf{r}_-} = -\epsilon \frac{4 \cdot f'|_{\mu(\mathbf{v})} \cdot d^{\boldsymbol{\lambda}}_{\mathbf{r}_+}}{(d^{\boldsymbol{\lambda}}_{\mathbf{r}_+} + d^{\boldsymbol{\lambda}}_{\mathbf{r}_-})^2} \boldsymbol{\Lambda} (\mathbf{v} - \mathbf{w}_{\mathbf{r}_-}), \tag{3.21}$$

with $\boldsymbol{\Lambda}$ being the diagonal matrix with entries $\lambda_1, \ldots, \lambda_{D_{\mathscr{V}}}$. Appropriate weighting factors $\boldsymbol{\lambda}$ can also be determined automatically via a stochastic gradient descent. In this way the GLVQ-learning rule (3.17) is augmented as follows

$$\Delta \lambda_k = -\epsilon_1 \cdot 2 \cdot f'|_{\mu(\mathbf{v})} \left(\frac{d^{\boldsymbol{\lambda}}_{\mathbf{r}_-}}{(d^{\boldsymbol{\lambda}}_{\mathbf{r}_+} + d^{\boldsymbol{\lambda}}_{\mathbf{r}_-})^2} (\mathbf{v} - \mathbf{w}_{\mathbf{r}_+})_k^2 - \frac{d^{\boldsymbol{\lambda}}_{\mathbf{r}_+}}{(d^{\boldsymbol{\lambda}}_{\mathbf{r}_+} + d^{\boldsymbol{\lambda}}_{\mathbf{r}_-})^2} (\mathbf{v} - \mathbf{w}_{\mathbf{r}_-})_k^2 \right), \tag{3.22}$$

with $\epsilon_1 \in (0,1)$ and $k = 1 \ldots D_{\mathscr{V}}$, followed by normalization to obtain $\|\boldsymbol{\lambda}\| = 1$. We term this generalization of the relevance learning vector quantization (RLVQ) (Bojer, Hammer, Schunk, & von Toschanowitz, 2001) and GLVQ generalized relevance learning vector quantization (GRLVQ) as introduced in Hammer and Villmann (2001a, 2002).

Obviously, the same idea could be applied to any gradient dynamics. We could, for example, minimize a different error function such as the Kullback–Leibler divergence of the distribution which is to be learned and the distribution which is implemented by the vector quantizer. This approach is not limited to supervised tasks. Unsupervised methods such as the NG (Martinetz et al., 1993), which obey a gradient dynamics, could be improved by application of weighting factors in order to obtain an adaptive metric. Furthermore, in unsupervised topographic mapping models like the SOM or NG one can try to learn the relevance of input dimensions subject to the requirement of topology preservation (Hammer & Villmann, 2001a,b).

3.4. Further approaches

Beside the above introduced methods further approaches of neural inspired algorithms are well known. In this context we have to refer to the family of fuzzy-clustering algorithms as for instance Fuzzy-SOM (Bezdek & Pal, 1993; Graepel, Burger, & Obermayer, 1998; Hasenjäger, Ritter, & Obermayer, 1999), Fuzzy-c-means (Bezdek, Pal, Hathaway, & Karayiannis, 1995; Gath & Geva, 1989) or other (neural) vector quantizers based on minimization of an energy function (Bishop, Svensén, & Williams, 1998; Buhmann & Kühnel, 1993; Linsker, 1989). For an overview of respective approaches we refer to Haykin (1994) (neural network approaches), Duda and Hart, (1973), and Schürmann (1996) (pattern classification).

4. Application in remote sensing image analysis

4.1. Low-dimensional data: LANDSAT TM multi-spectral images

4.1.1. Image description

LANDSAT TM satellite-based sensors produce images of the Earth in seven different spectral bands. The ground resolution is 30×30 m^2 for bands 1–5 and band 7. Band 6 (thermal band) is often dropped from analyses because of the lower spatial resolution. The LANDSAT TM bands were strategically determined for optimal detection and discrimination of vegetation, water, rock formations and cultural features within the limits of broad band multi-spectral imaging. The spectral information, associated with each pixel of a LANDSAT scene is represented by a vector $\mathbf{v} \in \mathscr{V} \subseteq \mathbb{R}^{D_{\mathscr{V}}}$ with $D_{\mathscr{V}} = 6$. The aim of any classification algorithm is to subdivide this data space into subsets of data points, with each subset corresponding to specific surface covers such as forest, industrial region, etc. The feature categories are specified by prototype data vectors (training spectra).

In the present contribution we consider a LANDSAT TM image from the Colorado area, USA.[6] In addition we have a manually generated label map (ground truth image) for the assessment of classification accuracy. All pixels of the Colorado image have associated ground truth labels, categorized into 14 classes. The classes include several

[6] Thanks to M. Augusteijn (Univerity of Colorado) for providing this image.

Table 1
Table of labels and classes occurring in the LANDSAT TM Colorado image

Class label	Ground cover type
a	Scotch pine
b	Douglas fir
c	Pine/fir
d	Mixed pine forest
e	Supple/prickle pine
f	Aspen/mixed pine forest
g	Without vegetation
h	Aspen
i	Water
j	Moist meadow
k	Bush land
l	Grass/pastureland
m	Dry meadow
n	Alpine vegetation
o	Misclassification

regions of different vegetation and soil, and can be used for supervised learning schemes like the GRLVQ (Table 1).

4.1.2. Analyses of LANDSAT TM imagery

A initial Grassberger–Procaccia analysis (Grassberger & Procaccia, 1983) yields $D^{\mathscr{GP}} \approx 3.1414$ for the intrinsic dimensionality (ID) of the Colorado image.

SOM analysis. The GSOM generates a $12 \times 7 \times 3$ lattice ($N_{\max} = 256$) in agreement with the Grassberger–Procaccia analysis ($D^{\mathscr{GP}} \approx 3.1414$), which corresponds to a $\tilde{P}$-value of 0.0095 indicating good topology preservation (Fig. 3).

One way to visualize the GSOM clusters of LANDSAT TM data is to use a SOM dimension $D_{\mathscr{A}} = 3$ (Gross & Seibert, 1993) and interpret the positions of the neurons $\mathbf{r}$ in the lattice $\mathscr{A}$ as vectors $\mathbf{c}(\mathbf{r}) = (r, g, b)$ in the RGB color space, where r, g, b are the intensities of the colors red, green and blue, respectively, computed as $c_l(\mathbf{r}) = [(r_l - 1)/(n_l - 1)]255$, for $l = 1, 2, 3$ (Gross & Seibert, 1993).

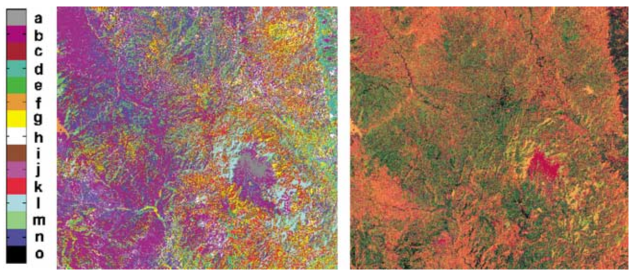

Fig. 3. Left: the reference classification for the Landsat TM Colorado image, provided by M. Augusteijn. Classed are keyed as shown on the color wedge on the left, and the corresponding surface units are described in Table 1. Right: cluster map of the Colorado image derived from all six bands by the $12 \times 7 \times 3$ GSOM-solution. Note that, due to the visualization scheme, which uses a continuous scale of RGB colors described in the text, this map does not show hard cluster boundaries. More similar colors indicate more similar spectral classes.

Such assignment of colors to winner neurons immediately yields a pseudo-color cluster map of the original image for visual interpretation. Since we are mapping the data clouds from a six-dimensional input space onto a three-dimensional color space dimensional conflicts may arise and the visual interpretation may fail. However, in the case of topologically faithful mapping this color representation, prepared using all six LANDSAT TM image bands, contains considerably more information than a customary color composite combining three TM bands (frequently bands 2, 3, and 3). (Gross & Seibert, 1993).

Application of GRLVQ. We trained the GRLVQ with 42 codebook vectors (three for each class) on 5% of the data set until convergence. The algorithm converged in <10 cycles for $\epsilon = 0.1$ and $\epsilon_1 = 0.01$. The GRLVQ produced a classification accuracy of about 90% on the training set as well as on the entire data set. The weighting factors resulting from GRLVQ analysis provide a ranking of the data dimensions

$$\boldsymbol{\lambda} = (0.1, 0.17, 0.27, 0.21, 0.26, 0.0). \quad (4.1)$$

Clearly, dimension 6 is ranked as least important with weighting factor close to 0. Dimension 1 is the second least important followed by dimensions 2 and 4. Dimensions 3 and 5 are of similarly high importance according to this ranking. If we prune dimensions 6, 1, and 2, an accuracy of 84% can still be achieved. Pruning dimension 4 brings the classification accuracy down to 50%, which is a very significant loss of information. This indicates that the intrinsic data dimension may not be as low as 2. These classification results are shown in Fig. 4 with misclassified pixels colored black and classes keyed by the color wedge.

Use of the scaled metric (3.18) during training (which is equivalent to having a new input space $\mathcal{V}_{\boldsymbol{\lambda}}$) of a GSOM results in a two-dimensional lattice structure the size of which is 11×10. This is in agreement with a corresponding Grassberger–Procaccia estimate of the intrinsic data dimension as $D_{\boldsymbol{\lambda}}^{\mathcal{GP}} \approx 2.261$ when this metric is applied. However, as we saw above and in Fig. 4, reducing the dimensionality of this Landsat TM image to 2 is not as simple as discarding the four input image planes that received the lowest weightings from the GRLVQ. Both the Grassberger–Procaccia estimate and the GSOM suggest the intrinsic dimension of 2 for $\mathcal{V}_{\boldsymbol{\lambda}}$ whereas the dimension suggested by GRLVQ is at least 4. Hence, the (scaled) data lie on a two-dimensional submanifold in $\mathcal{V}_{\boldsymbol{\lambda}}$. The distribution of weights for the two-dimensional lattice structure is visualized in Fig. 5 for each original (but scaled) input dimension. All input dimensions, except the fourth, seem to be correlated. The fourth dimension shows a clearly different distribution which causes a two-dimensional representation. The corresponding clustering, visualized in the two-dimensional $(r, g, 0)$ color space is depicted in Fig. 4. Based on visual inspection this cluster structure compares better with the reference classification in Fig. 4, upper left, than the GSOM clustering in Fig. 4. The scaling information provided to the GSOM by GRLVQ based on training labels seems to have improved the quality of the clustering.

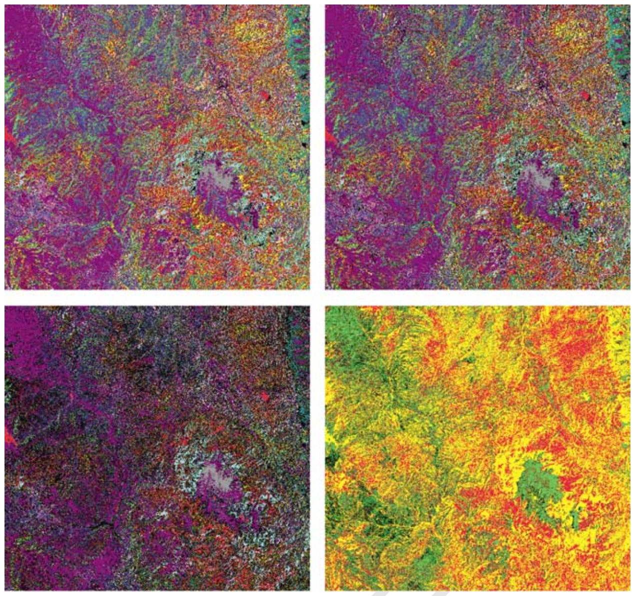

Fig. 4. GRLVQ results for the Landsat TM Colorado image. Upper left: GRLVQ without pruning. Upper right: GRLVQ with pruning of dimensions 1, 2, and 6. Lower left: GRLVQ with pruning dimensions 1, 2, 6, and 4. Lower right: Clustering of the Colorado image using GSOM with GRLVQ scaling. For the three supervised classifications, the same color wedge and class labels apply as in Fig. 3, left.

4.2. Hyperspectral data: the lunar crater volcanic field AVIRIS image

4.2.1. Image description

A visible–near infrared (0.4–2.5 μm), 224-band, 20 m/pixel AVIRIS image of the lunar crater volcanic

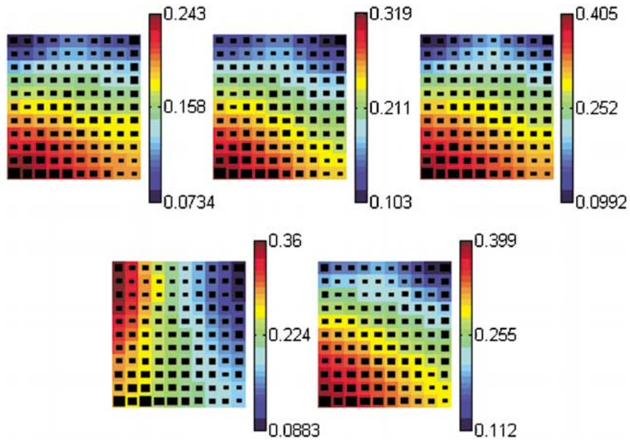

Fig. 5. The distribution of weight values in the two-dimensional 11×10 GSOM lattice for the non-vanishing input dimensions 1–5 as a result of GRLVQ-scaling of the Colorado image during GSOM training. Dimension 4 has significantly different weight distribution from all others, while in the rest of the dimensions the variations are less dramatic. (Please note that the scaling is different for each dimension.) For visualization the SOM-Toolbox provided by the Neural Network Group at Helsinki University of Technology was used.

field (LCVF), Nevada, USA, was analyzed in order to study SOM performance for high-dimensional remote sensing spectral imagery. (AVIRIS is the airborne visible-near infrared imaging spectrometer, developed at NASA/Jet Propulsion Laboratory. See http://makalu.jpl.nasa.gov for details on this sensor and on imaging spectroscopy). The LCVF is one of NASA's remote sensing test sites, where images are obtained regularly. A great amount of accumulated ground truth from comprehensive field studies (Arvidson et al., 1991) and research results from independent earlier work such as Farrand (1991) provide a detailed basis for the evaluation of the results presented here.

Fig. 6 shows a natural color composite of the LCVF with labels marking the locations of 23 different surface cover types of interest. This 10×12 km^2 area contains, among other materials, volcanic cinder cones (class A, reddest peaks) and weathered derivatives thereof such as ferric oxide rich soils (L, M, W), basalt flows of various ages (F, G, I), a dry lake divided into two halves of sandy (D) and clayey composition (E); a small rhyolitic outcrop (B); and some vegetation at the lower left corner (J), and along washes (C). Alluvial material (H), dry (N,O,P,U) and wet (Q,R,S,T) playa outwash with sediments of various clay contents as well as other sediments (V) in depressions of the mountain slopes, and basalt cobble stones strewn around the playa (K) form a challenging series of spectral signatures for pattern recognition (Merényi, 1998). A long, NW–SE trending scarp, straddled by the label G, borders the vegetated area. Since this color composite only contains information from three selected image bands (one red, one green, and one blue), many of the cover type variations remain undistinguished. They will become evident in the cluster and class maps below.

After atmospheric correction and removal of excessively noisy bands (saturated water bands and overlapping detector channels), 194 image bands remained from the original 224. These 194-dimensional spectra are the input patterns in the following analyses.

The spectral dimensionality of hyperspectral images is not well understood and it is an area of active research. While many believe that hyperspectral images are highly redundant because of band correlations, others maintain an opposite view. Few investigations exist into the ID of hyperspectral images. Linear methods such as PCA or determination of mixture model endmembers (Adams, Smith, & Gillespie, 1993; R.S. Inc, 1997) usually yield 3–8 'endmembers'. Optimally topology preserving maps, a TRN approach (Bruske & Merényi, 1999), finds the spectral ID of the LCVF AVIRIS image to be between 3 and 7 whereas the Grassberger–Procaccia estimate (Grassberger & Procaccia, 1983) for the same image is $D^{\mathcal{GP}} \approx 3.06$. These surprisingly low numbers, that increase with improved sensor performance (Green & Boardman, 2000), result from using statistical thresholds for the determination of what is 'relevant', regardless of application dependent meaning of the data.

The number of relevant components increases dramatically when specific goals are considered such as what cover classes should be separated. With an associative neural network, Pendock (1999) extracted 20 linear mixing

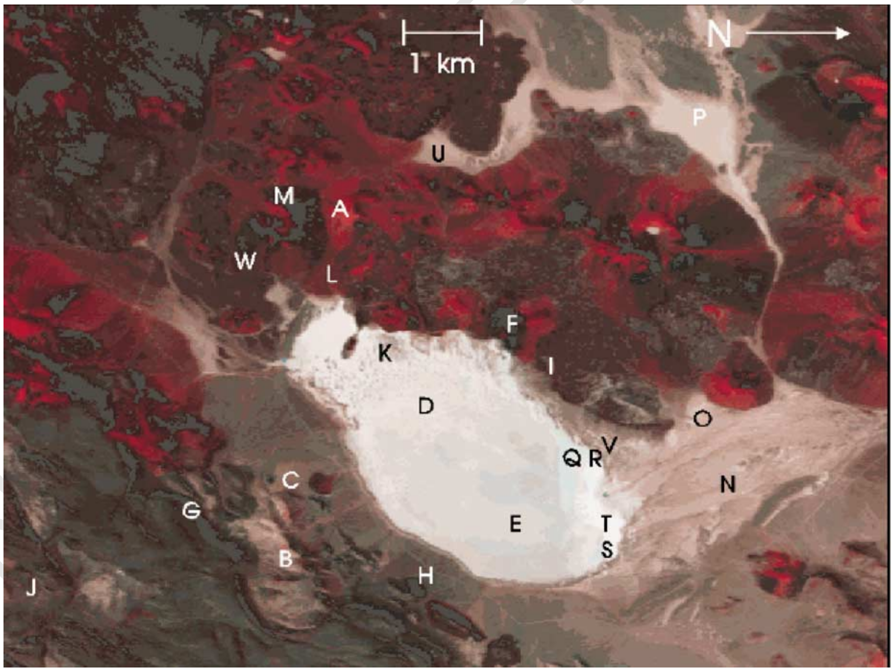

Fig. 6. The Lunar Crater Volcanic Field. RGB natural color composite from an AVIRIS, 1994 image. The full hyperspectral image comprises 224 image bands over the 0.4–2.5 μm wavelength range, 512 × 614 pixels, altogether 140 MB of data. Labels indicate 23 different cover types described in the text. The ground resolution is 20 m/pixel.

endmembers from a 50-band (2.0–2.5 μm) segment of an AVIRIS image of Cuprite, Nevada (another well known remote sensing test site), setting only a rather general surface texture criterium. Benediktsson et al. (1994) performed feature extraction on an AVIRIS geologic scene of Iceland, which resulted in 35 bands. They used an ANN (the same network that performed the classification itself) for decision boundary feature extraction (DBFE). The DBFE is claimed to preserve all features that are necessary to achieve the same accuracy by using all the original data dimensions, by the same classifier for predetermined classes. However, no comparison of classification accuracy was made using the full spectral dimension to support the DBFE claim. In their study a relatively low number of classes, 9, were of interest, and the goal was to find the number of features that describe those classes. Separation of a higher number of classes may require more features.

It is not clear how feature extraction should be done in order to preserve relevant information in hyperspectral images. Selection of 30 bands from our LCVF image by any of several methods leads to a loss of a number of the originally determined 23 cover classes. One example is shown in Fig. 8. Wavelet compression studies on an earlier image of the the same AVIRIS scene (Moon & Merényi, 1995) conclude that various schemes and compression rates affect different spectral classes differently, and none was found overall better than another, within 25–50% compressions (retaining 75–50% of the wavelet coefficients). In a study on simulated, 201-band spectral data, (Benediktsson et al., 1990) show slight accuracy increase across classifications on 20-, 40-, and 60-band subsets. However, that study is based on only two vegetation classes, the feature extraction is a progressive hierarchical subsampling of the spectral bands, and there is no comparison with using the full, 201-band case. Comparative studies using full spectral resolution and many classes are lacking, in general, because few methods can cope with such high-dimensional data technically, and the ones that are capable (such as minimum distance, parallel piped) often perform too poorly to merit consideration.

Undesirable loss of relevant information can result using any of these feature extraction approaches. In any case, finding an optimal feature extraction requires great preprocessing efforts just to tailor the data to available tools. An alternative is to develop capabilities to handle the full spectral information. Analysis of unreduced data is important for the establishment of benchmarks, exploration and novelty detection; as well as to allow for the distinction of significantly greater number of cover types than from traditional multi-spectral imagery (such as LANDSAT TM), according to the purpose of modern imaging spectrometers.

4.2.2. SOM analyses of hyperspectral imagery

A systematic supervised classification study was conducted on the LCVF image (Fig. 6), to simultaneously assess loss of information due to reduction of spectral dimensionality, and to compare performances of several traditional and an SOM-based hybrid ANN classifier. The 23 geologically relevant classes indicated in Fig. 6 represent a great variety of surface covers in terms of spatial extent, the similarity of spectral signatures (Merényi, 1998), and the number of available training samples. The full study, complete with evaluations of classification accuracies, is described in Merényi et al. (2001). Average spectral shapes of these 23 classes are also shown in Merényi (1998).

Fig. 7, top panel, shows the best classification, with 92% overall accuracy, produced by an SOM-hybrid ANN using all 194 spectral bands for input. This ANN first learns in an unsupervised mode, during which the input data are clustered in the hidden SOM layer. In this version the SOM uses the conscience mechanism of DeSieno (1988), which forces density matching (i.e. a magnification factor of 1) as pointed out in Section 3.2 and illustrated for this particular LCVF image and SOM mapping in Merenyi (2000). After the convergence of the SOM, the output layer is allowed to learn class labels via a Widrow–Hoff learning rule. The preformed clusters in the SOM greatly aid in accurate and sensitive classification, by helping prevent the learning of inconsistent class labels. Detailed description of this classifier is given in several previous scientific studies, which produced improved interpretation of high-dimensional spectral data compared to earlier analyses (Howell, Merényi, & Lebofsky, 1994; Merényi, Singer, & Miller, 1996; Merényi et al., 1997). Training samples for the supervised classifications were selected based on field knowledge. The SOM hidden layer was not evaluated and used for identification of spectral types (SOM clusters) prior to training sample determination. Fig. 7 reflects the geologist's view of the desirable segmentation.

In order to apply maximum likelihood and other covariance based classifiers, the number of spectral channels needed to be reduced to 30, since the maximum number of training spectra that could be identified for shape all classes was 31. Dimensionality reduction was performed in several ways, including PCA, equidistant subsampling, and band selection by a domain expert. Band selection by domain expert proved most favorable. Fig. 7, bottom panel, shows the maximum likelihood classification with 30 bands, which produced 51% accuracy. A number of classes (notably the ones with subtle spectral differences, such as N, Q, R, S, T, V, W) were entirely lost. Class K (basalt cobbles) disappeared from most of the edge of the playa, and only traces of B (rhyolitic outcrop) remained. Class G and F were greatly overestimated. Although the ANN classifier produced better results (not shown here) on the same 30-band reduced data set than the maximum likelihood, a marked drop in accuracy (to 75% from 92%) occurred compared to classification on the full data set. This emphasizes that accurate mapping of 'interesting', spatially small geologic units is possible from full hyperspectral information and with appropriate tools.

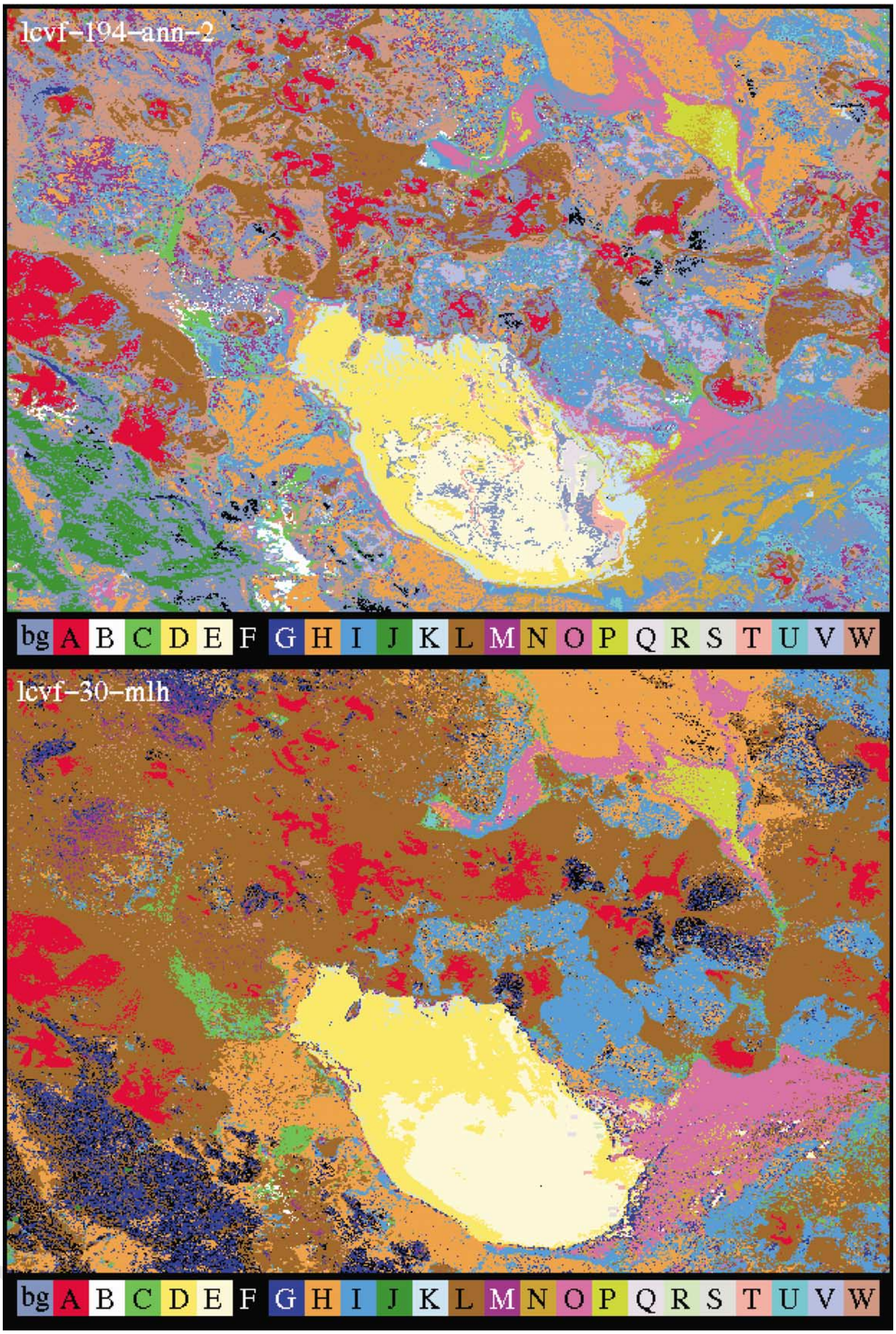

Fig. 7. Top: SOM-hybrid supervised ANN classification of the LCVF scene, using 194 image bands. The overall classification accuracy is 92% (Merényi et al., 2001). Bottom: maximum likelihood classification of the LCVF scene. 30, strategically selected bands were used due to the limited number of training samples for a number of classes. Considerable loss of class distinction occurred compared to the ANN classification, resulting in an overall accuracy of 51%. 'bg' stands for background (unclassified pixels).

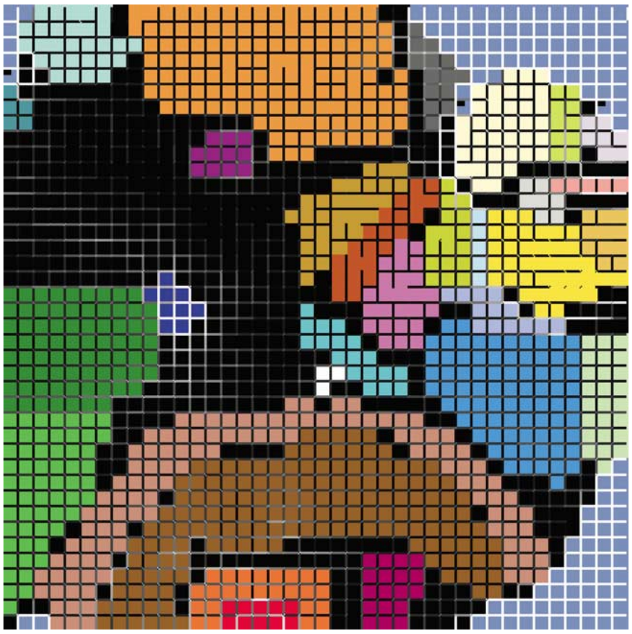

Fig. 8. Clusters identified in a 40 × 40 SOM. The SOM was trained on the entire 194-band LCVF image, using the DeSieno algorithm (DeSieno, 1988).

Discovery in Hyperspectral Images with a SOM. The previous section demonstrated the power of the SOM in helping to discriminate among a large number of pre-determined surface cover classes with subtle differences in the spectral patterns, using the full spectral resolution. It is even more interesting to examine the SOM's performance for the detection of clusters in high-dimensional data. Fig. 8 displays a 40 × 40 SOM trained with the conscience algorithm (DeSieno, 1988). The input data space was the entire 194-band LCVF image. Groups of neurons, altogether 32, that were found to be sensitized to groups of similar spectra in the 194-dimensional input data, are indicated by various colors. The boundaries of these clusters were determined by a somewhat modified version of the **U**-matrix method (Ultsch, 1992). The density matching property of this variant of the SOM facilitates proper mapping of spatially small clusters and therefore increases the possibility of discovery. Examples of discoveries are discussed later. Areas where no data points (spectra) were mapped are the grey corners with uniformly high fences, and are relatively small. The black background in the SOM lattice shows areas that have not been evaluated for cluster detection. The spatial locations of the image pixels mapped onto the groups of neurons in Fig. 8, are shown in the same colors in Fig. 9. Color coding for clusters that correspond to classes or subclasses of those in Fig. 7, top, is the same as in Fig. 7, to show similarities. Colors for additional groups were added.

The first observation is the striking correspondence between the supervised ANN class map in Fig. 7, top panel, and this clustering: the SOM detected all classes that were

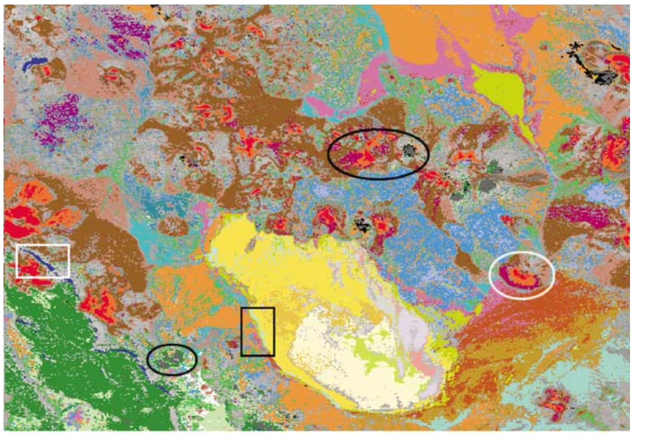

Fig. 9. The clusters from Fig. 8 remapped to the original spatial image, to show where the different spectral types originated from. The relatively large, light grey areas correspond to the black, unevaluated parts of the SOM in Fig. 8. Ovals and rectangles highlight examples of small classes with subtle spectral differences, discussed in the text. 32 clusters were detected, and found geologically meaningful, adding to the geologist's view reflected in Fig. 7, top panel.

known as meaningful geological units. The 'discovery' of classes B (rhyolitic outcrop, white), F (young basalt flows, dark grey and black, some shown in the black ovals), G (a different basalt, exposed along the scarp, dark blue, one segment outlined in the white rectangle), K (basalt cobbles, light blue, one segment shown in the black rectangle), and other spatially small classes such as the series of playa deposits (N, O, P, Q, R, S, T) is significant. This is the capability we need for sifting through high-volume, high-information-content data to alert for interesting, novel, or hard-to-find units. The second observation is that the SOM detected more, spatially coherent, clusters than the number of classes that we trained for in Fig. 7. The SOM's view of the data is more refined and more precise than that of the geologist's. For example, class A (in Fig. 7) is split here into a red (peak of cinder cones) and a dark orange (flanks of cinder cones) cluster that make geologic sense. The maroon cluster to the right of the red and dark orange clusters at the bottom of the SOM fills in some areas that remained unclassified (bg) in the ANN class map, in Fig. 7. An example is the arcuate feature at the base of the cinder cone in the white oval that apparently contains a material different enough to merit a separate spectral cluster. This material fills other areas, also unclassified in Fig. 7, consistently at the foot of cinder cones (another example is seen in the large black oval). Evaluation of further refinements are left to the reader. Evidence that the SOM mapping in Fig. 9 approximates an equiprobabilistic mapping (that the magnification factor for the SOM in Fig. 9 is close to 1), using DeSieno's algorithm, is presented in Merenyi (2000).

GSOM analysis of the LCVF hyperspectral image. As mentioned earlier, earlier investigations yielded an intrinsic spectral dimensionality of 3–7 for the LCVF data set

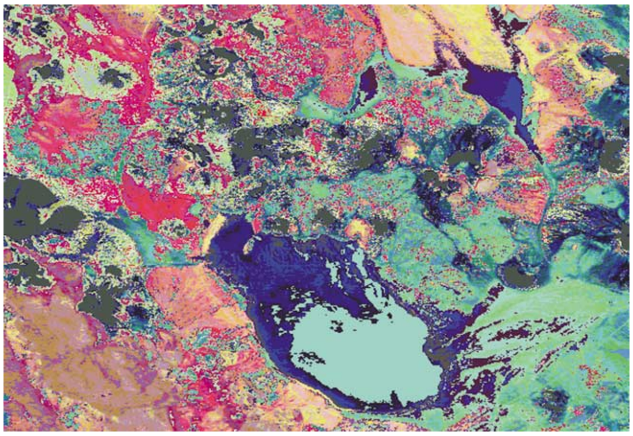

Fig. 10. GSOM-generated cluster map of the same 194 band hyperspectral image of the Lunar Crater Volcanic Field, Nevada, USA, as in Fig. 7. It shows groups similar to those in the supervised classification map in Fig. 7, top panel. Subtle spectral units such as B, G, K, Q, R, S, T were not separated, however, and for example, the yellow unit appears at both the locations of known young basalt deposits (class F, black, in Fig. 7, top panel) and at locations of alluvial deposits (class H, orange, in Fig. 7).

(Bruske & Merényi, 1999). The Grassberger–Procaccia estimate (Grassberger & Procaccia, 1983) $D_{\mathscr{A}}^{\mathscr{GP}} \approx 3.06$ corroborates the lower end of the above range, suggesting that the data are highly correlated, and therefore a drastic dimensionality reduction may be possible. However, a faithful topographic mapping is necessary to preserve the information contained in the hyperspectral image. In addition to the conscience algorithm explicit magnification control (Bauer et al., 1996) according to Eq. (3.10) and the growing SOM (GSOM) procedure, as extensions of the standard SOM, are suitable tools (Villmann, 2002). The GSOM produced a lattice of dimensions $8 \times 6 \times 6$ for the LCVF image, which represents a radical dimension reduction, and it is in agreement with the Grassberger–Procaccia analysis earlier. The resulting false color visualization of the spectral clusters is depicted in Fig. 10. It shows a segmentation similar to that in the supervised classification in Fig. 7, top panel. Closer examination reveals, however, that a number of the small classes with subtle spectral differences from others were lost (B, G, K, Q, R, S, T classes in Fig. 7, top panel). In addition, some confusion of classes can be observed. For example, the bright yellow cluster appears at locations of known young basalt outcrops (class F, black, in Fig. 7, and it also appears at locations of alluvial deposits (class H, orange in Fig. 7, top panel). This may inspire further investigation into the use of magnification control, and perhaps the growth criteria in the GSOM.

5. Conclusion

SOMs have been showing great promise for the analyses of remote sensing spectral images. With recent advances in remote sensor technology, very high-dimensional spectral data emerged and demand new and advanced approaches to cluster detection, visualization, and supervised classification. While standard SOMs produce good results, the high dimensionality and large amount of hyperspectral data call for very careful evaluation and control of the faithfulness of topological mapping performed by SOMs. Faithful topological mapping is required in order to avoid false interpretations of cluster maps created by an SOM. We summarized several new approaches developed in the past few years. Extensions to the standard Kohonen SOM, the growing SOM, magnification control, and GRLVQ were discussed along with the modified topographic product $\tilde{P}$. These ensure topology preservation through mathematical considerations. Their performance, and relationship to a former powerful SOM extension, the DeSieno conscience mechanism was discussed in the framework of case studies for both low-dimensional traditional multi-spectral, and very high-dimensional (hyperspectral) imagery. The Grassberger–Procaccia analysis served for an independent estimate of the determination of ID to benchmark ID estimation by GSOM and GRLVQ. While we show some excellent data clustering and classification, there remains certain discrepancy between theoretical considerations and application results, notably with regard to ID measures and the consequences of dimensionality reduction to classification accuracy. This will be targeted in future work. Finally, since it is outside the scope of this contribution, we want to point out that full scale investigations such the LCVF study also have to make heavy use of advanced image processing tools and user interfaces, to handle great volumes of data efficiently, and for effective graphics/visualization. References to such tools are made in the cited literature on data analyses.

Acknowledgements

E.M. has been supported by the Applied Information Systems Research Program of NASA, Office of Space Science, NAG59045 and NAG5-10432. Contributions by Dr William H. Farrand, providing field knowledge and data for the evaluation of the LCVF results, are gratefully acknowledged. Khoros (Khoral Research, Inc.) and NeuralWorks Professional (NeuralWare) packages were utilized in research software development.

References

Adams, J. B., Smith, M. O., & Gillespie, A. R. (1993). Imaging spectroscopy: Interpretation based on spectral mixture analysis. In C. Peters, & P. Englert (Eds.), *Remote geochemical analysis: Elemental and mineralogical composition* (pp. 145–166). New York: Cambridge University Press.

Arvidson, R. E., M.D.-B,, et al. (1991). Archiving and distribution of geologic remote sensing field experiment data. *EOS, Transactions of the American Geophysical Union*, *72*(17), 176.

Bauer, H.-U., Der, R., & Herrmann, M. (1996). Controlling the magnification factor of self-organizing feature maps. *Neural Computation*, *8*(4), 757–771.

Bauer, H.-U., & Pawelzik, K. R. (1992). Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, *3*(4), 570–579.

Bauer, H.-U., & Villmann, T. (1997). Growing a hypercubical output space in a self-organizing feature map. *IEEE Transactions on Neural Networks*, *8*(2), 218–226.

Benediktsson J. A., Sveinsson, J. R., et al (1994). Classification of very-high-dimensional data with geological applications. *Proceedings of MAC Europe 91* (pp. 13–18), Germany: Lenggries.

Benediktsson, J. A., Swain, P. H., et al (1990). Classification of very high dimensional data using neural networks. *IGARSS'90 10th Annual International Geoscience and Remote Sensing Symposium* (Vol. 2, p. 1269).

Bezdek, J., & Pal, N. R (1993). Fuzzification of the self-organizing feature map: Will it work? *Proceedings of the SPIE—The International Society for Optical Engineering*, *2061* (p. 142–162).

Bezdek, J. C., Pal, N. R., Hathaway, R. J., & Karayiannis, N. B (1995). Some new competitive learning schemes. *Proceedings of the SPIE—The International Society for Optical Engineering 2492* (pt. 1) (pp. 538–549) (Applications and Science of Artificial Neural Networks Conference, 17–21 April 1995, Orlando, FL, SPIE).

Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998). GTM: the generative topographic mapping. *Neural Computation*, *10*, 215–234.

Bojer, T., Hammer, B., Schunk, D., & von Toschanowitz, K. T (2001). Relevance determination in learning vector quantization. *Proceedings of European Symposium on Artificial Neural Networks (ESANN'2001)* (pp. 271–276). Brussels, Belgium: D facto publications.

Bruske, J., & Merényi, E (1999). Estimating the intrinsic dimensionality of hyperspectral images, *Proceedings of European Symposium on Artificial Neural Networks (ESANN'99)* (pp. 105–110), Brussels, Belgium: D facto publications.

Buhmann, J., & Kühnel, H. (1993). Vector quantization with complexity costs. *IEEE Transactions on Information Theory*, *39*, 1133–1145.

Campbell, J. (1996). *Introduction to remote sensing*. USA: The Guilford Press.

Clark, R. N. (1999). Spectroscopy of rocks and minerals, and principles of spectroscopy. In A. Rencz (Ed.), *Manual of remote sensing*. New York: Wiley.

DeSieno, D (1988). Adding a conscience to competitive learning, *Proceedings of ICNN'88, International Conference on Neural Networks* (pp. 117–124) Piscataway, NJ: IEEE Service Center.

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Erwin, E., Obermayer, K., & Schulten, K. (1992). Self-organizing maps: ordering, convergence properties and energy functions. *Biological Cybernetics*, *67*(1), 47–55.

Farrand, W. H (1991). *VIS/NIR reflectance spectroscopy of tuff rings and tuff cones*. PhD Thesis. University of Arizona.

Fritzke, B. (1995). In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *A growing neural gas network learns topologies* (*Vol. 7*). *Advances in neural information processing systems*, Cambridge, MA: MIT Press.

Gath, I., & Geva, A. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*, 773–781.

Goppert, J., & Rosenstiel, W. (1997). The continuous interpolating self-organizing map. *Neural Processing Letters*, *5*(3), 185–192.

Graepel, T., Burger, M., & Obermayer, K. (1998). Self-organizing maps: generalizations and new optimization techniques. *Neurocomputing*, *21*(1–3), 173–190.

Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, *9*, 189–208.

Green, R. O (1996). *Summaries of the Sixth Annual JPL Airborne Geoscience Workshop*. AVIRIS Workshop, Pasadena, CA, March 4–6.

Green, R. O., & Boardman, J (2000). Exploration of the relationship between information content and signal-to-noise ratio and spatial resolution. *Proceedings of the Ninth AVIRIS Earth Science and Applications Workshop*, Pasadena, CA.

Gross, M. H., & Seibert, F. (1993). Visualization of multidimensional image data sets using a neural network. *Visual Computer*, *10*, 145–159.

Hammer, B., & Villmann, T. (2001a). Estimating relevant input dimensions for self-organizing algorithms. In N. Allinson, H. Yin, L. Allinson, & J. Slack (Eds.), *Advances in self-organising maps* (pp. 173–180). London: Springer.

Hammer, B., & Villmann, T (2001b). Input pruning for neural gas architectures, *Proceedings of European Symposium on Artificial Neural Networks (ESANN'2001)* (pp. 283–288), Brussels, Belgium: D facto publications.

Hammer, B., & Villmann, T. (2002). Generalized relevance learning vector quantization. *Neural Networks*, *15*(8/9), 1059–1068.

Hasenjäger, M., Ritter, H., & Obermayer, K. (1999). Active data selection for fuzzy topographic mapping of proximities. In G. Brewka, R. Der, S. Gottwald, & A. Schierwagen (Eds.), *Fuzzy-neuro-systems'99 (FNS'99)* (pp. 93–103). *Proceedings of the FNS-Workshop*, Leipzig: Leipziger Universitätsverlag.

Haykin, S. (1994). *Neural networks—A comprehensive foundation*. New York: IEEE Press.

Howell, E. S., Merényi, E., & Lebofsky, L. A. (1994). Classification of asteroid spectra using a neural network. *Journal of Geophysics Research*, *99*(10), 847–865.

Kaski, S (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. *Proceedings of IJCNN'98, International Joint Conference on Neural Networks* (Vol. 1, pp. 413–418), Piscataway, NJ: IEEE Service Center.

Kaski, S., & Sinkkonen, J. (1999). *Metrics induced by maximizing mutual information. Publications in computer and information science, report A55*, Espoo, Finland: Helsinki University of Technology.

Kohonen, T. (1995). *Self-organizing maps* (*Vol. 30*). *Springer series in information sciences*, Berlin: Springer, Second extended edition 1997.

Kohonen, T. (1999). Comparison of SOM point densities based on different criteria. *Neural Computation*, *11*(8), 212–234.

Kohonen, T., Kaski, S., Lappalainen, H., & Saljärvi, H (1997). The adaptive-subspace self-organizing map (assom), *International Workshop on Self-Organizing Maps (WSOM'97)* (pp. 191–196), Helsinki.

Linsker, R. (1989). How to generate maps by maximizing the mutual information between input and output signals. *Neural Computation*, *1*, 402–411.

Martinetz, T. M., Berkovich, S. G., & Schulten, K. J. (1993). 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, *4*(4), 558–569.

Martinetz, T., & Schulten, K. (1994). Topology representing networks. *Neural Networks*, *7*(3), 507–522.

Merényi, E (1998). Self-organizing ANNs for planetary surface composition research. *Proceedings of European Symposium on Artificial Neural Networks (ESANN'98)* (pp. 197–202), Brussels, Belgium: D facto publications.

Merényi, E (1999). The challenges in spectral image analysis: An introduction and review of ANN approaches. *Proceedings of European Symposium on Artificial Neural Networks (ESANN'99)* (pp. 93–98), Brussels, Belgium: D facto publications.

Merenyi, E. (2000). 'Precision mining' of high-dimensional patterns with self-organizing maps: Interpretation of hyperspectral images. In P. Sincak, & J. Vascak (Eds.), *Quo vadis computational intelligence? New trends and approaches in computational intelligence* (*Vol. 54*). *Studies in fuzziness and soft computing*, Physica-Verlag.

Merényi, E., Howell, E. S., et al. (1997). Prediction of water in asteroids from spectral data shortward of 3 microns. *ICARUS*, *129*(10), 421–439.

Merényi, E., Farrand, W. H., et al (2001). Efficient geologic mapping from hyperspectral images with artificial neural networks classification: a comparison to conventional tools. *IEEE TGARS* (in preparation).

Merényi, E., Singer, R. B., & Miller, J. S. (1996). Mapping of spectral variations on the surface of mars from high spectral resolution telescopic images. *ICARUS*, *124*(10), 280–295.

Meyering, A., & Ritter, H (1992). Learning 3D-shape-perception with local linear maps. *Proceedings of IJCNN'92* (pp. 432–436).

Moon, T., & Merényi, E (1995). Classification of hyperspectral images using wavelet transforms and neural networks. *Proceedings of the Annual SPIE Conference* (p. 2569), San Diego, CA.

Pendock, N (1999). A simple associative neural network for producing spatially homogeneous spectral abundance interpretations of hyperspectral imagery. *Proceedings of European Symposium on Artificial Neural Networks (ESANN'99)* (pp. 99–104), Brussels, Belgium: D facto publications.

Pregenzer, M., Pfurtscheller, G., & Flotzinger, D. (1996). Automated feature selection with a distinction sensitive learning vector quantizer. *Neurocomputing*, *11*(1), 19–29.

R. S. Inc., ENVI v.3 User's Guide (1997).

Richards, J., & Jia, X. (1999). *Remote sensing digital image analysis*. Berlin: Springer, Third, revised and enlarged edition.

Ritter, H. (1993). Parametrized self-organizing maps. In S. Gielen, & B. Kappen (Eds.), *Proceedings of the ICANN'93 International Conference on Artificial Neural Networks* (pp. 568–575). London: Springer.

Ritter, H., & Schulten, K. (1986). On the stationary state of Kohonen's self-organizing sensory mapping. *Biological Cybernetics*, *54*, 99–106.

Sato, A. S., & Yamada, K. (1995). Generalized learning vector quantization. In G. Tesauro, D. Touretzky, & L. Leen (Eds.), (*Vol. 7*) (pp. 423–429). *Advances in neural information processing systems*, Cambridge, MA: MIT Press.

Schürmann, J. (1996). *Pattern classification*. New York: Wiley.

Somervuo, P., & Kohonen, T. (1999). Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, *10*(2), 151–159.

Ultsch, A. (1992). Knowledge acquisition with self-organizing neural networks. In I. Aleksander, & J. Taylor (Eds.), (*Vol. I*) (pp. 735–738). *Artificial neural networks, 2*, Amsterdam: North-Holland.

Villmann, T. (2000). Controlling strategies for the magnification factor in the neural gas network. *Neural Network World*, *10*(4), 739–750.

Villmann, T. (2002). Neural maps for faithful data modelling in medicine—state of the art and exemplary applications. *Neurocomputing*, *48*(1–4), 229–250.

Villmann, T., Der, R., Herrmann, M., & Martinetz, T. (1997). Topology preservation in self-organizing feature maps: exact definition and measurement. *IEEE Transactions on Neural Networks*, *8*(2), 256–266.

Villmann, T., & Herrmann, M (1998). Magnification control in neural maps. *Proceedings of European Symposium on Artificial Neural Networks (ESANN'98)* (pp. 191–196), Brussels, Belgium: D facto publications.

van Hulle, M. M. (2000). *Faithful representations and topographic maps from distortion- to information-based self-organization*. London: Wiley.