Class imaging of hyperspectral satellite remote sensing data using FLSOM

T. Villmann[1], F.-M. Schleif[1], E. Merenyi[2], M. Strickert[3], and B. Hammer[4]
[1] University Leipzig, [2] Rice University Houston,
[3] IPK Gatersleben, [4] University of Technology Clausthal
email: thomas.villmann@medizin.uni-leipzig.de

Keywords: SOM, fuzzy classification, visualization

***Abstract*— We propose an extension of the self-organizing map for supervised fuzzy classification learning, whereby uncertain (fuzzy) class information is also allowed for training data. The method is able to detect class similarities, which can be used for data vizualization. Applying a special functional metric, derived from of the L_p norms, we show the application of the method for classification and visualization of hyper-spectral data in satellite image remote sensing image analysis.

1 Introduction

The self-organizing map (SOM) introduced by T. KOHONEN constitutes one of the most popular data mining and visualization methods for processing of high-dimensional and complex data [12]. It is based on principles of prototype based unsupervised vector quantization whereby a topological grid structure with neighborhood cooperativeness is installed on the set of prototypes, usually chosen as a rectangular two-dimensional lattice. However, other arrangements are possible. Under certain conditions the SOM prototype adjustment (learning) generates a model which allows a nonlinear mapping of the given data set onto a the low-dimensional regular lattice in a topology-preserving fashion [12] for easy data analysis and interpretation [20].

During the last years, several extensions of the basic SOM have been established to make the approach more flexible and to assess the quality of the generated model [23]. These are related to adaptive lattice structures, to the processing of structured data and to handling of labeled data by supervised learning. Thereby, the handling of appropriate, problem dependent data metrics becomes also more and more important (metric adaptation) [8].

Although the learning scheme of the SOM is quite simple, its mathematical foundation is non-trivial. Most theoretical results are only valid for special cases, whereas more general approaches become intractable. In particular, it is a blemish that the usual SOM does not follow a gradient descent on any cost function [5] such that the final state is not well defined. This problem can be solved by a small modification of the original learning scheme, as it was shown by HESKES [11].

This modification offers new possibilities for supervised learning using SOMs, i.e. for active utilization of data labels during training. Several extensions of the unsupervised SOM were presented for processing supervised classification tasks ranging from simple post-labeling, the well-known counter-propagation network to combine SOMs and multilayer perceptrons or fuzzy decision schemes [10, 12]. However, all these methods have in common that the prototype learning of the underlying SOM is not influenced by the subsequent classification learning, and, hence, the prototypes are not adjusted in dependence of the classification task. The FuzzySOM proposed by P. VUORIMAA imposes a supervised learning vector quantization scheme for the SOM-prototypes (LVQ) on an unsupervised trained usual SOM to learn a classification task [24]. The subsequent LVQ learning rule does not minimizes the classification error. Thus both parts of FuzzySOM are based on heuristics and, therefore, have not well-determined optimization goals. Moreover, the topographic mapping learned during the unsupervised SOM phase may be violated by the classification learning, because neighborhood cooperativeness is not integrated in LVQ.

Here we propose the utilization of the cost function according to HESKES in combination with a misclassification penalization term as the new cost function for supervised SOM. Thereby fuzzy class memberships of training data are allowed, i.e. uncertain class information is can be used. Minimizing of the proposed cost function by gradient descent leads to the fuzzy labeled SOM (FLSOM). Like the usual SOM, the FLSOM generates a topology preserving data mapping onto the SOM grid for faithful training conditions and proper data. Moreover, each prototype is equipped with a fuzzy label vector describing its probabilistic or possibilistic class membership. Using the topology preservation property of the SOM one can derive class similarities by investigation of the spatial distribution of the class membership within the lattice environment, which finally allows similarity preserving visualization of the classification by multi-dimensional scaling (MDS) of the label vectors.

The power of the FLSOM is demonstrated for hyperspectral image analysis of satellite remote sensing spectral data. In this context a special data similarity mea-

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

sure is used based on the Minkowski-norm. This so-called (parametric) functional norm takes the spatial correlations within each spectrum into account. We further integrate the idea of metric adaptation into the above FLSOM scheme which improves classification and gives a task dependent filtering of the spectra. This is done by optimizing of the functional norm as a gradient descent with respect to the norm parameters.

2 Fuzzy-Labeled SOM

2.1 Basic SOM with cost function

Originally, the SOM is an unsupervised learning of topographic vector quantization such that data are mapped onto a regular grid of nodes (neurons). Assume data $\mathbf{v} \in \mathcal{V} \subseteq \mathbb{R}^{D_\mathcal{V}}$ are given distributed according to an underlying distribution $P(\mathcal{V})$. A SOM is determined by a set A of neurons $\mathbf{r}$ equipped with weight vectors/prototypes $\mathbf{w_r} \in \mathbb{R}^{D_\mathcal{V}}$ and arranged on a lattice structure which determines the neighborhood or topological relation $N(\mathbf{r}, \mathbf{r}')$, the discrete grid distance, between neurons $\mathbf{r}$ and $\mathbf{r}'$. Denote the set of prototypes by $\mathbf{W} = \{\mathbf{w_r}\}_{\mathbf{r}\in A}$. In the SOM variant according to HESKES [11], the mapping description of a trained SOM defines a function

$$\Psi_{\mathcal{V}\to A} : \mathbf{v} \mapsto s(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{r}\in A} de(\mathbf{v}, \mathbf{r}). \tag{1}$$

with *local (data) errors*

$$de(\mathbf{v}, \mathbf{r}) = \sum_{\mathbf{r}'\in A} h_\sigma(\mathbf{r}, \mathbf{r}')\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \tag{2}$$

and

$$h_\sigma(\mathbf{r}, \mathbf{r}') = \exp\left(-\frac{N(\mathbf{r}, \mathbf{r}')}{\sigma}\right) \tag{3}$$

determines the neighborhood cooperation with range $\sigma > 0$. $\xi(\mathbf{v}, \mathbf{w})$ is an appropriate distance measure, usually the standard Euclidean norm

$$\xi(\mathbf{v}, \mathbf{w_r}) = \|\mathbf{v} - \mathbf{w_r}\|^2 = (\mathbf{v} - \mathbf{w_r})^2. \tag{4}$$

However, here we assume $\xi(\mathbf{v}, \mathbf{w})$ to be arbitrary supposing that it is a differentiable and symmetric function which measures some data similarity. In this formulation, an input stimulus is mapped onto that position $\mathbf{r}$ of the SOM, where the distance $\xi(\mathbf{v}, \mathbf{w_r})$ is minimum, whereby the average over all neurons according to the neighborhood is taken. We refer to this neuron $s(\mathbf{v})$ as the winner.

During the adaptation process a sequence of data points $\mathbf{v} \in \mathcal{V}$ is presented to the map representative for the data distribution $P(\mathcal{V})$. Each time the currently most proximate neuron $s(\mathbf{v})$ according to (1) is determined. All weights within the neighborhood of this neuron are adapted by

$$\triangle\mathbf{w_r} = -\epsilon h_\sigma(\mathbf{r}, s(\mathbf{v})) \frac{\partial\xi(\mathbf{v}, \mathbf{w_r})}{\partial\mathbf{w_r}} \tag{5}$$

with learning rate $\epsilon > 0$. This adaptation follows the stochastic gradient descent of the cost function

$$E_{\mathrm{SOM}} = \frac{1}{2C(\sigma)} \int P(\mathbf{v}) \sum_{\mathbf{r}} \delta_{\mathbf{r}}^{s(\mathbf{v})} \cdot de(\mathbf{v}, \mathbf{r})\, d\mathbf{v} \tag{6}$$

were $C(\sigma)$ is a constant which we will drop in the following, and $\delta_{\mathbf{r}}^{\mathbf{r}'}$ is the usual Kronecker symbol checking the identity of $\mathbf{r}$ and $\mathbf{r}'$.

One main aspect of SOMs is the visualization ability of the resulting map due to its topological structure. Under certain conditions the resulting non-linear projection $\Psi_{\mathcal{V}\to A}$ generates a continuous mapping from the data space $\mathcal{V}$ onto the grid structure A [21]. This mapping can mathematically be interpreted as an approximation of the principal curve or its higher-dimensional equivalents [9]. Thus, as pointed out above, similar data points are projected on prototypes which are neighbored in the grid space A. Further, prototypes neighbored in the lattice space should code similar data properties, i.e. their weight vectors should be close together in the data space according to the metric ξ. This property of SOMs is called topology preserving (or topographic) mapping realizing the mathematical concept of continuity. For a detailed consideration of this topic we refer to [21].

2.2 Integrating fuzzy classification into SOM

We now integrate the label (class) information into the learning scheme of SOM to allow supervised learning. This is done in such a way that the prototype adjustment is depending on both the data distribution as well as the label information. Assume training point $\mathbf{v}$ is equipped with a label vector $\mathbf{x} \in [0, 1]^C$ describing the class information of C classes, whereby the component x_i of $\mathbf{x}$ determines the probabilistic/possibilistic assignment of $\mathbf{v}$ to class i for $i = 1, \ldots, C$. Hence, we can interpret the label vector as probabilistic or possibilistic fuzzy class memberships. In case of probabilistic labeled data we have the constraint $\sum_{i=1}^{C} x_i = 1$ and for crisp labeled data the additional condition $x_i \in \{0, 1\}$ holds. Accordingly, we add to each prototype vector $\mathbf{w_r}$ of the map a label vector $\mathbf{y_r} \in [0, 1]^C$ which determines the amount of neuron $\mathbf{r}$ assigned to the respective classes. The new cost function to be minimized contains two terms: the unsupervised part E_{SOM} and a new one E_{FL} describing the classification accuracy

$$E_{\mathrm{FLSOM}} = (1 - \beta) E_{\mathrm{SOM}} + \beta E_{\mathrm{FL}} \tag{7}$$

where $\beta \in [0, 1]$ is a balance factor to determine the influence of the goal of clustering the data set and the goal of achieving a correct labeling. One can simply choose $\beta = 0.5$, for example. For classification accuracy we choose

$$E_{\mathrm{FL}} = \frac{1}{2} \int P(\mathbf{v}) \sum_{\mathbf{r}} ce(\mathbf{v}, \mathbf{r})\, d\mathbf{v} \tag{8}$$

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

with *local, weighted classification errors*

$$ce\left(\mathbf{v},\mathbf{r}\right)=g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)\left(\mathbf{x}-\mathbf{y_r}\right)^2 \tag{9}$$

and $g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)$ is a Gaussian kernel describing a neighborhood range in the data space:

$$g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)=\exp\left(-\frac{\xi\left(\mathbf{v},\mathbf{w_r}\right)}{2\gamma^2}\right). \tag{10}$$

This choice is based on the assumption that data points close to the prototype determine the corresponding label if the underlying classification is sufficiently smooth. Note that $g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)$ depends on the prototype locations, such that E_{FL} is influenced by both $\mathbf{w_r}$ and $\mathbf{y_r}$, and, hence, $\frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{w_r}}$ contributes to the usual SOM-learning by

$$\frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{w_r}}=-\frac{1}{4\gamma^2}\int P\left(\mathbf{v}\right)\cdot ce\left(\mathbf{v},\mathbf{r}\right)\cdot\frac{\partial \xi\left(\mathbf{v},\mathbf{w_r}\right)}{\partial \mathbf{w_r}}d\mathbf{v} \tag{11}$$

The label update is independent from the first term E_{SOM} such that it simply becomes

$$\frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{y_r}}=-\int P\left(\mathbf{v}\right)\cdot g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)\cdot\left(\mathbf{x}-\mathbf{y_r}\right)d\mathbf{v} \tag{12}$$

It can be interpreted as a weighted average of the data fuzzy labels of those data close to the associated prototypes.

As mentioned above, unsupervised SOMs generate a topographic mapping from the data space onto the prototype grid under specific conditions. If the classes are consistently determined with respect to the varying data, one can expect for supervised topographic FLSOMs that the labels become ordered within the grid structure of the prototype lattice. In this case the topological order of the prototypes should be transferred to the topological order of prototype labels such that we have a smooth change of the fuzzy class label vectors between neighbored grid positions. This is the consequence of following fact: the neighborhood function $h_{\sigma}\left(\mathbf{r},\mathbf{s}\right)$ of the usual SOM learning (5) forces the topological ordering of the prototypes. In FLSOM, this ordering is further influenced by the weighted classification error $ce\left(\mathbf{v},\mathbf{r}\right)$, which contains the data space neighborhood $g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)$, eq. (11). Hence, the prototype ordering contains information of both data density and class distribution, whereby for high value β the latter term becomes dominant. Otherwise, the data space neighborhood $g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)$ also triggers the label learning (12), which is, of course, also dependent on the underlying learned prototype distribution and ordering. Thus, a consistent ordering of the labels is obtained in FLSOM.

Hence, the evaluation of the similarities between the prototype label vectors yields suggestions for the similarity of classes, i.e. similar classes are represented by prototypes in a local spatial area of the SOM lattice. In case of overlapping class distributions the topographic processing leads to prototypes with unclear decision, located between prototypes with clear vote. Further, if classes are not distinguish-able, there will exist prototypes responsive to those data which have class label vectors containing approximately the same degree of class membership for the respective classes.

2.3 Functional norm and metric adaptation

In usual SOMs the data similarity measure $\xi\left(\mathbf{v},\mathbf{w_r}\right)$ is usually chosen to be the squared Euclidean metric. However, depending on task, this choice may be not optimum. Therefore, other measure are also of interest, for example TANIMOTO's distance or correlation measures in taxonomy or medicine/biology. Now we consider a parametrized differentiable distance measure $\xi^{\lambda}(\mathbf{v},\mathbf{w})$ with a parameter vector $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_M)$ with $\lambda_i\geq 0$ and normalization $\sum_i\lambda_i=1$. The idea of metric adaptation or relevance learning is to optimize the parameters λ_i with respect to the classification task using a gradient descent [7]. Formal derivation yields

$$\frac{\partial E_{\mathrm{FLSOM}}}{\partial \lambda_l}=(1-\beta)\frac{\partial E_{\mathrm{SOM}}}{\partial \lambda_l}+\beta\frac{\partial E_{\mathrm{FL}}}{\partial \lambda_l} \tag{13}$$

We obtain

$$\frac{\partial E_{\mathrm{SOM}}}{\partial \lambda_l}=\frac{1}{2}\int P(\mathbf{v})\sum_{\mathbf{r}}\delta_{\mathbf{r}}^{s(\mathbf{v})}\cdot\frac{\partial de\left(\mathbf{v},\mathbf{r}\right)}{\partial \lambda_l}d\mathbf{v} \tag{14}$$

with

$$\frac{\partial de\left(\mathbf{v},\mathbf{r}\right)}{\partial \lambda_l}=\sum_{\mathbf{r}'}h_{\sigma}(\mathbf{r},\mathbf{r}')\cdot\frac{\partial \xi^{\lambda}(\mathbf{v},\mathbf{w_r})}{\partial \lambda_l} \tag{15}$$

and

$$\frac{\partial E_{\mathrm{FL}}}{\partial \lambda_l}=-\frac{1}{4\gamma^2}\int P(\mathbf{v})\sum_{\mathbf{r}}ce\left(\mathbf{v},\mathbf{r}\right)\cdot\frac{\partial \xi^{\lambda}(\mathbf{v},\mathbf{w_r})}{\partial \lambda_l}d\mathbf{v} \tag{16}$$

for the respective parameter adaptation.

In case of $\xi^{\lambda}(\mathbf{v},\mathbf{w})$ being the scaled Euclidean metric

$$\xi^{\lambda}(\mathbf{v},\mathbf{w})=\sum_i\lambda_i(v_i-w_i)^2\,, \tag{17}$$

relevance learning ranks the input dimensions i according to their relevance for the classification task at hand. Thus, the corresponding learning rule for the relevance parameters becomes

$$\begin{aligned}\triangle\lambda_l &= -\epsilon_{\lambda}\frac{1-\beta}{2}\sum_{\mathbf{r}}h_{\sigma}(s(\mathbf{v}),\mathbf{r})\cdot(v_l-(w_{\mathbf{r}})_l)^2 \\ &\quad +\epsilon_{\lambda}\frac{\beta}{4\gamma^2}\sum_{\mathbf{r}}g_{\gamma}(\mathbf{v},\mathbf{w_r})(v_l-(w_{\mathbf{r}})_l)^2(\mathbf{x}-\mathbf{y_r})^2\end{aligned} \tag{18}$$

(subscript l denoting the component l of a vector) with learning rate $\epsilon_{\lambda}>0$. This update is followed by normalization to ensure $\lambda_i\geq 0$ and $\sum_i\lambda_i=1$.

The Euclidean metric takes the input dimensions as independent. However, for functional data like spectra or

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

time series, the data dimensions correspond to frequencies or time points, respectively, and, therefore, they are spatially correlated. LEE&VERLEYSEN proposed a *functional norm* which takes these correlations implicitly into account [13]. For a vectorial representation $\mathbf{v}$ of a function we assume that between neighbored data dimensions there is a constant frequency or time difference. Then, the functional p-norm is defined as

$$\mathcal{L}_p^{fc}(\mathbf{v}) = \left(\sum_{k=1}^{D} \left(A_k(\mathbf{v}) + B_k(\mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{19}$$

with the terms

$$A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2} |v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k| + |v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \tag{20}$$

and

$$B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2} |v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k| + |v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \tag{21}$$

For $p = 2$ it induces the quadratic functional metric $\delta(\mathbf{v}, \mathbf{w_r}) = \left(\mathcal{L}_2^{fc}(\mathbf{v} - \mathbf{w_r}) \right)^2$. In analogy to the scaled Euclidean metric we introduce the *scaled* quadratic functional metric

$$\delta_{\boldsymbol{\lambda}}(\mathbf{v}, \mathbf{w_r}) = \left(\mathcal{L}_2^{fc}(\boldsymbol{\Lambda}(\mathbf{v} - \mathbf{w_r})) \right)^2 \tag{22}$$

with $\boldsymbol{\Lambda}$ being a diagonal matrix with entries $\Lambda_{ii} = \lambda_i$ and $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. The derivative $\frac{\partial \delta_{\boldsymbol{\lambda}}(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}}$ determines the learning rules and is obtained for the of the kth dimension as:

$$\left(\frac{\partial \delta_{\boldsymbol{\lambda}}(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}} \right)_k = (2 - U_{k-1} - U_{k+1})(V_{k-1} + V_{k+1}) \triangle_k \tag{23}$$

with

$$U_{k-1} = \begin{cases} 0 & \text{if } 0 \leq \triangle_k \triangle_{k-1} \\ \left(\frac{\lambda_{k-1} \triangle_{k-1}}{\lambda_k |\triangle_k| + \lambda_{k-1} |\triangle_{k-1}|} \right)^2 & \text{if } 0 > \triangle_k \triangle_{k-1} \end{cases}$$

$$U_{k+1} = \begin{cases} 0 & \text{if } 0 \leq \triangle_k \triangle_{k+1} \\ \left(\frac{\lambda_{k+1} \triangle_{k+1}}{\lambda_k |\triangle_k| + \lambda_{k+1} |\triangle_{k+1}|} \right)^2 & \text{if } 0 > \triangle_k \triangle_{k+1} \end{cases}$$

$$V_{k-1} = \begin{cases} 1\lambda_k & \text{if } 0 \leq \triangle_k \triangle_{k-1} \\ \frac{\lambda_k |\triangle_k|}{\lambda_k |\triangle_k| + \lambda_{k-1} |\triangle_{k-1}|} & \text{if } 0 > \triangle_k \triangle_{k-1} \end{cases}$$

$$V_{k+1} = \begin{cases} 1\lambda_k & \text{if } 0 \leq \triangle_k \triangle_{k+1} \\ \frac{\lambda_k |\triangle_k|}{\lambda_k |\triangle_k| + \lambda_{k+1} |\triangle_{k+1}|} & \text{if } 0 > \triangle_k \triangle_{k+1} \end{cases}$$

and $\triangle_k = v_k - w_k$.

Further, the derivative of $\delta_{\boldsymbol{\lambda}}(\mathbf{v}, \mathbf{w_r})$ with respect to the metric parameters λ_k, which have to be plugged into the gradient formulae (14) and (16) for metric adaptation, are

$$\frac{\partial \delta_{\boldsymbol{\lambda}}(\mathbf{v}, \mathbf{w_r})}{\partial \lambda_k} = \begin{cases} c_k |v_k| & \text{if } 0 \leq v_{k-1} v_k \\ z(k, k-1) |v_k| & \text{if } 0 > v_{k-1} v_k \end{cases} + \begin{cases} c_k |v_k| & \text{if } 0 \leq v_{k+1} v_k \\ z(k, k+1) |v_k| & \text{if } 0 > v_{k+1} v_k \end{cases}$$

with $c_j = A_j(\boldsymbol{\lambda}\mathbf{v}) + B_j(\boldsymbol{\lambda}\mathbf{v})$ and

$$z(k, j) = \frac{\lambda_k^2 c_k v_k^2 - c_j v_j^2 \lambda_j^2 + 2\lambda_k c_k |v_k| |v_j| \lambda_j}{(\lambda_k |v_k| + |v_j| \lambda_j)^2}$$

.

3 HiT-MDS-2 for class label visualization

Now we turn to use the learned class relations for visualization. As described in Sec.2.2, the fuzzy label vectors $\mathbf{y_r}$ of the FLSOM reflect class similarities. For visualization of class membership of data we suggest a color representation. Thereby, we make use of the learned class similarities. Thus we look for a similarity based representation. For this purpose, we map the prototype label vectors $\mathbf{y_r}$ onto color vectors $\mathbf{c_r}$ here chosen as RGB-vectors representing the color intensities for the colors red, green and blue. Yet, other color space representations are possible.

Similarity preserving mapping can be obtained by several approaches, for example by three-dimensional SOM, MDS or local linear embedding. Here we apply an advanced MDS scheme called *HiT-MDS-2*, which is more robust than usual MDS [19]. We briefly explain this method in the following.

Generally, MDS refers to the optimization of N point locations $\mathbf{t}_i = \left(t_i^1, \ldots, t_i^{\check{D}} \right) \in \mathbb{R}^{\check{D}}$ in a *target space* in such a way that their distance relationships faithfully reflect those of the associated *original data vectors* $\mathbf{o}_i \in O \subseteq \mathbb{R}^D$ [3]. Obviously, in case of dimension reduction with $D > \check{D}$ such optimization will need to find a compromise solution. Let $\delta_{i,k} = \delta(\mathbf{o}_i, \mathbf{o}_k)$ be the similarity (distance) measure in the original data space O. Further, let $d_{i,k} = d(\mathbf{t}_i, \mathbf{t}_k)$ be the distance in the lower-dimensional target space $\mathbb{R}^{\check{D}}$. If distances are Euclidean then the minimum of the canonical point-embedding stress function $J = \sum_{i<k} (d_{i,k} - \delta_{i,k})^2 = \min$ yields target configurations which are equivalent to the linear projections of principal component analysis (PCA). Although the benefit over PCA is more flexibility in the choice of distance measures, like many other metric MDS approaches, minimization of the respective stress function suffers from the presence of local minima. One avoidable reason for local minima is a too stringent formulation of the stress function. In most metric approaches, reconstructed distances are forced onto

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

a pre-defined line, such as the one with unit slope for the canonical stress, in the corresponding $\delta_{i,k}$-vs.-$d_{i,k}$ Shepard plot. In contrast to that, HiT-MDS-2 maximizes the Pearson correlation $r \in [-1;1]$

$$\begin{aligned} r &= \frac{\sum_{q<k}(\delta_{q,k}-\mu_\delta)\cdot(d_{q,k}-\mu_d)}{\sqrt{\sum_{q<k}(\delta_{q,k}-\mu_\delta)^2 \cdot \sum_{q<k}(d_{q,k}-\mu_d)^2}} \\ &= \frac{\mathcal{B}}{\sqrt{\mathcal{C}\cdot\mathcal{D}}} \end{aligned}$$

between entries of the source distances and the corresponding target space distances by minimizing negative Fisher's Z′:

$$J_{Z'} = -\frac{1}{2}\log\left(\frac{a+r}{a-r}\right) \stackrel{!}{=} \min, \qquad a = 1+\epsilon \quad (24)$$

Thereby, μ_δ and μ_d are the averaged distances in the data and target space, respectively. Locations of points $\mathbf{t}_i$ in target space are obtained by iterative gradient descent $\triangle t_i^k = -\varepsilon\frac{\partial J_{Z'}}{\partial t_i^k}$ of step size ε on the stress function $J_{Z'}$ using the chain rule:

$$\frac{\partial J_{Z'}}{\partial r} = -\frac{a}{r^2-a^2} \quad (25)$$

$$\frac{\partial r}{\partial d_{i,j}} = \frac{(\delta_{i,j}-\mu_\delta)\cdot\mathcal{D}-(d_{i,j}-\mu_d)\cdot\mathcal{B}}{\mathcal{D}\cdot\sqrt{\mathcal{C}\cdot\mathcal{D}}} \quad (26)$$

$$\frac{\partial d_{i,j}}{\partial t_i^k} = \frac{2\left(t_i^k-t_j^k\right)}{d_{i,j}} \quad \text{(for Euclidean target space)} \quad (27)$$

These equations yield a substantially improved convergence over the old formulation of HiT-MDS proposed earlier [19]. HiT-MDS-2 makes use of two non-critical parameters, the learning rate $\varepsilon = 0.1$ and the extra Z′ infinity-preventer $\epsilon = 0.001$, which in Fisher's original formula would be zero. While, for plotting purposes, target distances are usually Euclidean, however, input distances can be mere dissimilarities such as correlation.

A further remark is due to optimized computation of MDS embedding in case of incremental adaptation: The computational cost can be dramatically reduced using the fact that for each iteration only a few distances are really changed [19].

4 Application

We applied the FLSOM for classification of hyper spectral images in satellite remote sensing image analysis. Airborne and satellite-borne remote sensing spectral images consist of an array of multi-dimensional vectors (spectra) assigned to particular spatial regions (pixel locations) reflecting the response of a spectral sensor at various wavelengths. A spectrum 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. The number of applications has dramatically increased in the past years with the advent of imaging spectrometers such as AVIRIS of NASA/JPL. [22]

Imaging spectrometers sample a spectral window contiguously with very narrow, $10-20\,\mathrm{nm}$ bandpasses [18]. Depending on the wavelength resolution and the width of the wavelength window the dimensionality of the spectra can as high as several hundred [6].

Spectral images can be formally described as a matrix $\mathbf{S} = \mathbf{v}^{(x,y)}$, where $\mathbf{v}^{(x,y)} \in \mathbb{R}^{D_\mathcal{V}}$ is the vector of spectral information associated with pixel location (x,y). The elements $v_i^{(x,y)}$, $i = 1 \ldots D_\mathcal{V}$ of spectrum $\mathbf{v}^{(x,y)}$ reflect the responses of a spectral sensor at a suite of wavelengths [4]. The spectrum is a characteristic fingerprint pattern that identifies the surface material within the area defined by pixel (x,y). The individual 2-dimensional image $\mathbf{S}_i = v_i{}^{(x,y)}$ at wavelength i is called the ith image band. The data space $\mathcal{V}$ spanned by Visible-Near Infrared reflectance spectra is $[0-noise, U+noise]^{D_\mathcal{V}} \subseteq \mathbb{R}^{D_\mathcal{V}}$ 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 $\mathcal{P}(\mathcal{V})$ 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 [16].

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 structures, 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 the above difficulties motivate research into advanced and novel approaches.

A Visible-Near Infrared ($0.4-2.5$ μm), 224-band, 20 m/pixel AVIRIS image of the Lunar Crater Volcanic Field (LCVF), Nevada, U.S.A., was analyzed in order to study FLSOM performance for high-dimensional remote sensing spectral imagery. (AVIRIS is the Airborne Visible-Near Infrared Imaging Spectrometer, developed at NASA/Jet Propulsion Laboratory.

Figure 1 shows a natural color composite of the LCVF with labels marking the locations of 23 different surface cover types of interest. This $10 \times 12\,\mathrm{km}^2$ 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 (see in [16]). 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. After atmospheric correction and removal of excessively noisy bands (saturated water bands and overlapping detector channels), 194 image bands remained from the original 224, i.e. $D_{\mathcal{V}} = 194$. The 23 geologically relevant classes indicated in Figure 1 represent a great variety of surface covers in terms of spatial extent, the similarity of spectral signatures [16], and the number of available training samples $N = 931$.

Figure 1, middle panel, visualizes the best classification, with 92% overall image accuracy, produced by an SOM-MLP-hybrid network [15]. This network first learns in an unsupervised mode the hidden SOM layer. After the SOM convergence, the output layer is allowed to learn class labels via a *Widrow-Hoff* learning rule. Training samples for the supervised classifications were selected based on field knowledge.

In a second study a generalized learning vector quantization scheme (GRLVQ, [7]) was applied [14]. The overall number of prototypes was chosen as 115, i.e. 5 prototypes for each class. The achieved accuracy for the available training samples is 97.0%, whereby a scaled Euclidean metric was applied togehter with relevance learning for inproved performance.

To be comparable to the latter approach, the FLSOM lattice structure was chosen as 112 prototypes in a 16×7 grid. The grid edge length ratio was determined using the growing SOM [2]. The final balancing parameter was $\beta = 0.85$. The topography of the FLSOM is quite good giving a topographic product value nearby zero [1]. Small violations were detected by the topographic function [21]. The FLSOM accuracy for training samples (majority vote) is obtained as 95.3% for the Euclidean metric and 95.7% for the functional metric. The reduced accuracy in comparison to GRLVQ is due to the β-value, which means a non-vanishing term of unsupervised learning in the cost function. However, further increasing of β would lead to lost of the neighborhood regularization, which is needed for detecting class similarities based on topography properties (see Sec.2.2). In fact, the resulted distribution of the label vectors on the grid shows a clear ordering and smooth transitions, whereby classes with similar meaning are grouped together (Figure 2). This result can be evaluated by validity measures assessing cluster partitions in fuzzy clustering. There exist a broad range of indices [17]. Roughly speaking, most of the measures take mainly the compactness and the separability of clusters for judgement into account. Taking the class label distribution as 'cluster distribution', we can adapt them to the specific task of assessing the quality of the label distribution in the FLSOM-grid.

Here we used the validity index V_m provided in [17]:

$$V_m = J_m\left(A, \mathbf{Y}\right) - K_m\left(A, \mathbf{Y}\right) \tag{28}$$

with

$$J_m\left(A, \mathbf{Y}\right) = \sum_{\mathbf{r} \in A} \sum_{i=1}^{C} \left(y_{\mathbf{r}}^i\right)^m \cdot d_A\left(\mathbf{r}, \mathbf{c}_i\right) \tag{29}$$

is the cost function of *fuzzy-c-means* assessing the compactness of the clusters. Thereby, $\mathbf{y}_{\mathbf{r}} = \left(y_{\mathbf{r}}^1, \ldots, y_{\mathbf{r}}^C\right)$ are the label vectors of the prototypes, and $\mathbf{c}_i$ is the center location of class i. It is defined as the lattice location of the label with the highest class assignment

$$\mathbf{c}_i = \underset{\mathbf{r} \in A}{\operatorname{argmax}} \left(y_{\mathbf{r}}^i\right) \tag{30}$$

d_A is the Euclidean distance taken the grid indices as locations.

The second term $S_m\left(A, \mathbf{Y}\right)$ in (28) is the *separation index*

$$S_m\left(A, \mathbf{Y}\right) = \frac{1}{C \cdot \#A} \sum_{\mathbf{r} \in A} \sum_{i=1}^{C} \left(y_{\mathbf{r}}^i\right)^m \cdot d_A\left(\mathbf{r}, \bar{\mathbf{c}}\right) \tag{31}$$

judging the separability of clusters with $\bar{\mathbf{c}} = \frac{1}{C} \sum_{i=1}^{C} \mathbf{c}$ being here the mean of all grid locations of class centers $\mathbf{c}_i$.

The obtained label distribution in the application (see Figure 2) shows a clearly improved validity V_m index compared to labeling achieved by usual SOM-learning with subsequent post-labeling. Both terms J_m and S_m are independently optimized such that both covered features, compactness and separability, show improved performance for FLSOM leading to better interpretability.

Comparing the SOM-MLP-hybrid ANN visualization with the visualization obtained by FLSOM with subsequent HiT-MDS-2 color mapping (Figure 1, bottom), the first observation is the striking correspondence. Yet, the optimized coloring by class label mapping using HiT-MDS-2, which takes the class similarities detected by FLSOM into account (see Figure 2), leads to a more smoothed visualization, whereby similar material are represented by similar colors. For example, prototypes responsible for similar materials (wet playa - classes Q, R, S, T; alluvium - classes C, H, M; dry wash - classes N, O , P) are in small grid areas and the respective class label show a continuous transition, which generates similar color representation in visualization. Evaluation of further refinements are left to future work.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

5 Conclusion

We propose an extension of SOMs for fuzzy classification. The approach allows the detection of class similarities. For this purpose, the neighborhood cooperativnes of SOMs is used and transferred also to the supervised part of training for classification learning. The achieved similarity information can be used for better visualization of classification results. We demonstrate the method for hyper-spectral data classification and visualization in satellite remote sensing image analysis.

Acknowledgements

EM has been partially supported by NASA AISRP grant NNG05GA94G.

References

[1] H.-U. Bauer, K. Pawelzik, and T. Geisel. A topographic product for the optimization of self-organizing feature maps. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 1141–1147. Morgan Kaufmann, San Mateo, CA, 1992.

[2] H.-U. Bauer and T. Villmann. Growing a Hypercubical Output Space in a Self–Organizing Feature Map. *IEEE Transactions on Neural Networks*, 8(2):218–226, 1997.

[3] A. Buja, D. Swayne, M. Littman, N. Dean, and H. Hofmann. Interactive Data Visualization with Multidimensional Scaling. Report, University of Pennsylvania, 2004. http://www.ggobi.org/.

[4] J. Campbell. *Introduction to Remote Sensing*. The Guilford Press, U.S.A., 1996.

[5] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: Ordering, convergence properties and energy functions. *Biol. Cyb.*, 67(1):47–55, 1992.

[6] R. O. Green. *Summaries of the 6th Annual JPL Airborne Geoscience Workshop*. Pasadena, CA, March 4–6 1996.

[7] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.

[8] B. Hammer and T. Villmann. Classification using non-standard metrics. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2005)*, pages 303–316, Brussels, Belgium, 2005. d-side publications.

[9] T. Hastie and W. Stuetzle. Principal curves. *J. Am. Stat. Assn.*, 84:502–516, 1989.

[10] R. Hecht-Nielsen. Counterprogagation networks. *Appl. Opt.*, 26(23):4979–4984, December 1987.

[11] T. Heskes. Energy functions for self-organizing maps. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 303–316. Elsevier, Amsterdam, 1999.

[12] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[13] J. Lee and M. Verleysen. Generalization of the l_p norm for time series and its application to self-organizing maps. In M. Cottrell, editor, *Proc. of Workshop on Self-Organizing Maps (WSOM) 2005*, pages 733–740, Paris, Sorbonne, 2005.

[14] M. Mendenhall. *A Neural Relevance Model for Feature Extraction from Hyperspectral Images, and its Application in the Wavelet Domain*. PhD thesis, Rice University, Houston, TX, August 2006.

[15] E. Merenyi. "precision mining" of high-dimensional patterns with self-organizing maps: Interpretation of hyperspectral images. In P. Sincak and J. Vascak, editors, *Quo Vadis Computational Intelligence? New Trends and Approaches in Computational Intelligence (Studies in Fuzziness and Soft Computing, Vol. 54*. Physica-Verlag, ?, 2000.

[16] E. Merényi. Self-organizing ANNs for planetary surface composition research. In *Proc. Of European Symposium on Artificial Neural Networks (ESANN'98)*, pages 197–202, Brussels, Belgium, 1998. D facto publications.

[17] N. Pal and J. Bezdek. On the cluster validity for the Fuzzy *c*-means model. *IEEE Transactions on Fuzzy Systems*, 3(3):370–379, 1995.

[18] J. Richards and X. Jia. *Remote Sensing Digital Image Analysis*. Springer-Verlag, Berlin, Heidelberg, New York, third, revised and enlarged edition edition, 1999.

[19] M. Strickert, S. Teichmann, N. Sreenivasulu, and U. Seiffert. High-Throughput Multi-Dimensional Scaling (HiT-MDS) for cDNA-array expression data. In W. Duch et al., editor, *Artificial Neural Networks: Biological Inspirations, Part I, LNCS 3696*, pages 625–634. Springer, 2005. http://hitmds.webhop.net/.

[20] J. Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 3(7):123–456, 1999.

[21] T. Villmann, R. Der, M. Herrmann, and T. Martinetz. Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266, 1997.

[22] T. Villmann, E. Merényi, and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3-4):389–403, 2003.

[23] T. Villmann, U. Seiffert, and A. Wismüller. Theory and applications of neural maps. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks 2004*, pages 25–38. d-side publications, 2004.

[24] P. Vuorimaa. Fuzzy self-organizing map. *Fuzzy Sets and Systems*, 66(2):223–231, Sept 1994.

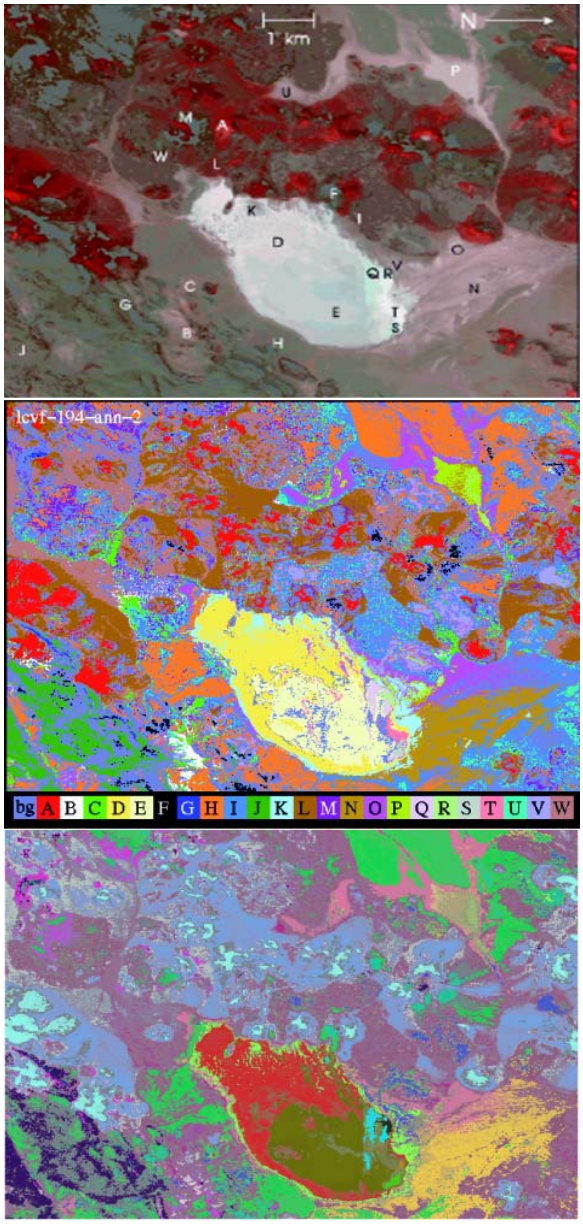

Figure 1: **top** - natural color composite of the LCVF with labels marking the locations of 23 different surface cover types (see text); **middle** - color representation of the classification result of the SOM-MLP-hyprid network; **bottom** - similarity based color representation of the classification result of the FLSOM approach using HiT-MDS-2 mapping.

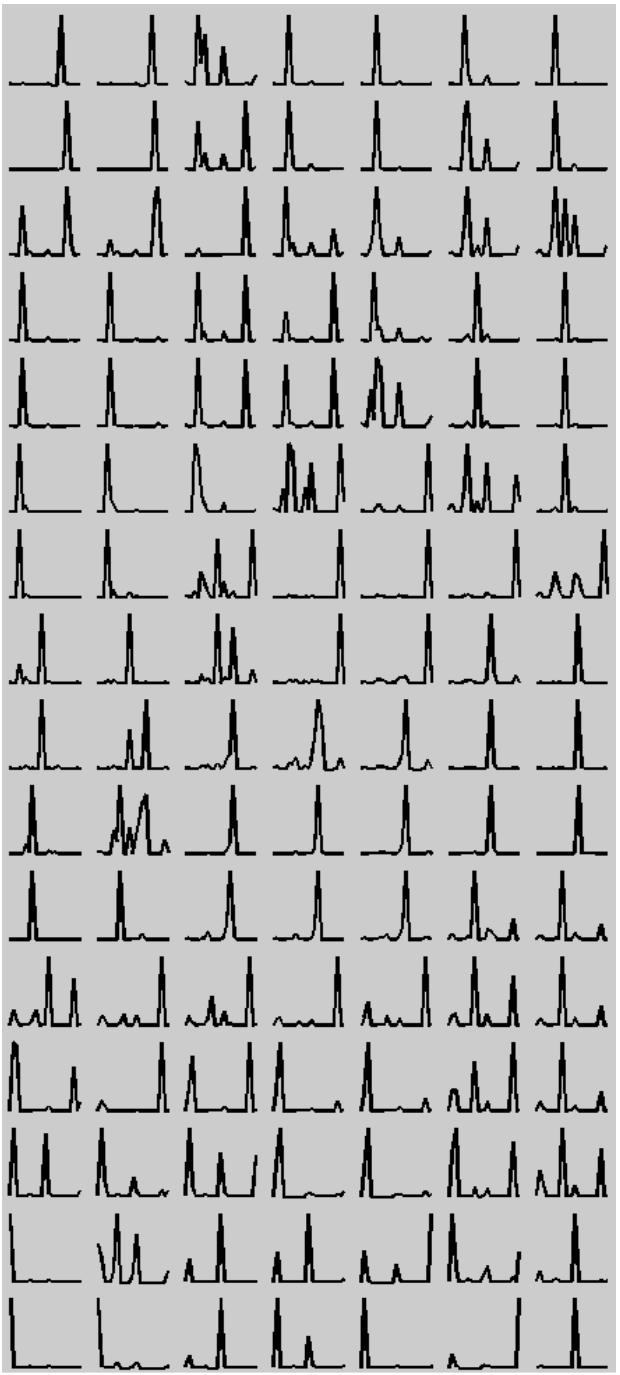

Figure 2: Visualization of the class label distribution within the FLSOM lattice. A clear ordering can be observed, which is the convergence of the class similarity learning.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de