Fuzzy Labeled Self-Organizing Map with Label-Adjusted Prototypes

Thomas Villmann[1,⋆], Udo Seiffert[2], Frank-Michael Schleif[3,4], Cornelia Brüß[2], Tina Geweniger[4,5], and Barbara Hammer[6]

[1] University Leipzig, Medical Department
[2] IPK Gatersleben, Pattern Recognition Group
[3] BRUKER Daltonik Leipzig, Numerical Toolbox Group
[4] University Leipzig, Institute of Computer Science
[5] University of Applied Science Mittweida, Computer Science
[6] Clausthal University of Technology, Institute of Computer Science

Abstract. We extend the self-organizing map (SOM) in the form as proposed by Heskes to a supervised fuzzy classification method. On the one hand, this leads to a robust classifier where efficient learning with fuzzy labeled or partially contradictory data is possible. On the other hand, the integration of labeling into the location of prototypes in a SOM leads to a visualization of those parts of the data relevant for the classification.

1 Introduction

The self-organizing map (SOM) constitutes one of the most popular data mining and visualization methods, mapping a given possibly high-dimensional data set nonlinearly onto a low-dimensional regular lattice in a topology-preserving fashion [10]. It can be taken as an adaptive *unsupervised* learning scheme for prototype based vector quantization with the additional feature of topographic mapping. Several methods exist to extend the SOM model for supervised classification tasks. These approaches range from simple post-labeling to the well-known counterpropagation network [10],[7],[8]. However, all these methods have in common that the locations of the prototypes in the data space remain unchanged by the subsequent determination of the prototype labels.

In the following we will propose an extension of the SOM such that it can be used as a prototype based classification approach. Thereby, the position of the prototypes is explicitly influenced by the classification task. In this way a combination of statistical and class properties triggers the prototype and the related label learning. The learning rules for prototypes as well as prototype labels are obtained from a cost function which is a combination of a classification error and the energy function of the SOM according to the formulation introduced by Heskes [9]. Thereby, the class information of the data may be fuzzy. The resulting map allows a visualization of the classification process by

⋆ Corresponding author.

F. Schwenker and S. Marinai (Eds.): ANNPR 2006, LNAI 4087, pp. 46–56, 2006.

means of the properties of topology preserving mapping of SOMs, which leads to a better understanding of the classification scheme. Further, metric adaptation, as known from learning vector quantization [4],[3], can be easily incorporated into this approach to improve its flexibility.

2 The Self-Organizing Map

As mentioned above, SOMs can be taken as unsupervised learning of topographic vector quantization with a topological structure (grid) within the set of prototypes (codebook vectors). Thereby, roughly speaking, topology preservation means that similar data points are mapped onto identical or neighbored grid locations (prototypes), see Fig 1. An exact mathematical definition is given in [12]. Successful tools for assessing this map property are the topographic function and the topographic product [12],[1].

There exists a wide range of applications in pattern recognition ranging from spectral image processing to bioinformatics. The mathematics behind the original model as proposed by Kohonen is rather complicated, particularly due to the lack of an underlying cost function for continuous data distributions. However, Heskes proposed a minor variant of the original algorithm which usually leads to the same results as the original SOM but for which a cost function can be established [9]. We will base our model on this formulation.

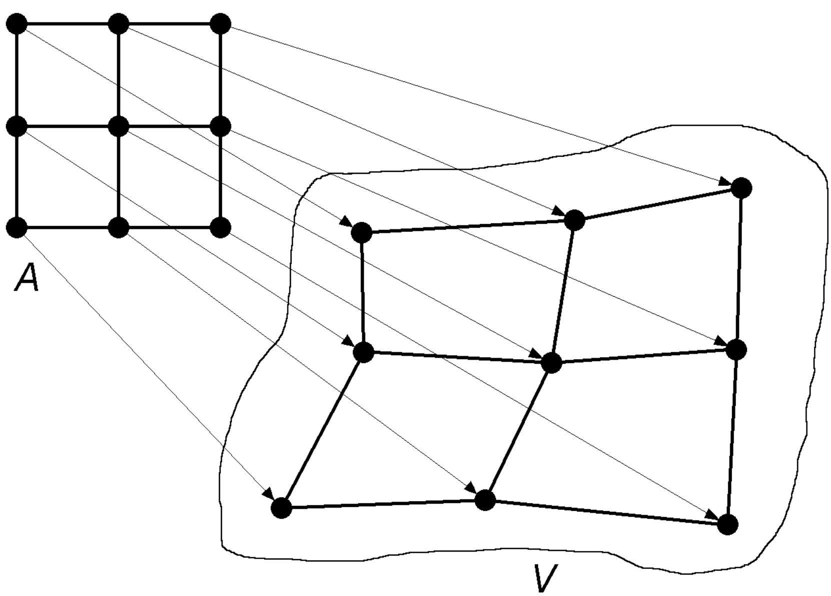

Fig. 1. Illustration of tropographic mapping by SOMs. A continuous change in the input space $\mathcal{V}$ leads to a representation by weight vectors, the neurons of which are neighbored in the grid space A.

Assume data $\mathbf{v} \in \mathcal{V}$ are given distributed according to an underlying distribution $P(\mathcal{V})$. A SOM is determined by a set A of neurons/prototypes $\mathbf{r}$ equipped with weight vectors $\mathbf{w_r} \in \mathbb{R}^d$ and arranged on a lattice structure which determines the neighborhood relation $N(\mathbf{r}, \mathbf{r}')$ of neuron $\mathbf{r}$ and $\mathbf{r}'$. Denote the set of neurons by $\mathbf{W} = \{\mathbf{w_r}\}_{\mathbf{r}\in A}$. The mapping description of a trained SOM defines a function

$$\Psi_{\mathcal{V}\to A} : \mathbf{v} \mapsto s(\mathbf{v}) = \underset{\mathbf{r}\in A}{\operatorname{argmin}} \sum_{\mathbf{r}'\in A} h_\sigma(\mathbf{r}, \mathbf{r}')\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) . \tag{1}$$

where

$$h_\sigma(\mathbf{r}, \mathbf{r}') = \exp\left(\frac{N(\mathbf{r}, \mathbf{r}')}{\sigma}\right) \tag{2}$$

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

$$\xi(\mathbf{v}, \mathbf{w_r}) = \|\mathbf{v} - \mathbf{w_r}\| = (\mathbf{v} - \mathbf{w_r})^2 . \tag{3}$$

However, here we chose $\xi(\mathbf{v}, \mathbf{w})$ to be arbitrary supposing that it is a differentiable and symmetric function which measures some 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

$$\triangle\mathbf{w_r} = -\epsilon h_\sigma(\mathbf{r}, s(\mathbf{v})) \frac{\partial\xi(\mathbf{v}, \mathbf{w_r})}{\partial\mathbf{w_r}} \tag{4}$$

with learning rate $\epsilon > 0$. This adaptation follows a stochastic gradient descent of the cost function introduced by Heskes [9]:

$$E_{\text{SOM}} = \frac{1}{2C(\sigma)} \int P(\mathbf{v}) \sum_{\mathbf{r}} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}'} h_\sigma(\mathbf{r}, \mathbf{r}')\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) d\mathbf{v} \tag{5}$$

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. This mapping can mathematically be interpreted as an approximation of the principal curve or its higher-dimensional equivalents [6]. 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 shoud 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 [12].

3 Fuzzy Labeled SOM (FLSOM)

SOM is a well-established model for nonlinear data visualization which, due to its above mentioned topology preserving properties, can also serve as an adequate preprocessing step for data completion, classification or interpolation. For high-dimensional data sets, however, the result is often suboptimal if no further information about the data is present. In such cases, a default model and metric such as the Euclidean metric often accounts for the fact that only general properties, but not necessarily the parts relevant for the task at hand are represented. Often, auxiliary data in the form of (possibly partial or contradictory) labels are available. In this case, SOM can be used for a preprocessing step in classification by means of posterior labeling. Here, we seek for an integration of the label information such that the prototype locations are determined with respect to the auxiliary data. On the one hand, this improves the classification result if we are interested in supervised classification. On the other hand, information relevant for the classification can be visualized by means of the underlying SOM topology adapted towards the labeling.

Assume training point $\mathbf{v}$ is equipped with a label vector $\mathbf{x} \in [0,1]^{N(c)}$ whereby the component x_i of $\mathbf{x}$ determines the assignment of $\mathbf{v}$ to class i for $i = 1, \ldots, N(c)$. Hence, we can interprete the label vector as probabilistic or possibilistic fuzzy class memberships. Accordingly, we enlarge each prototype vector $\mathbf{w}_\mathbf{r}$ of the map by a label vector $\mathbf{y}_\mathbf{r} \in [0,1]^{N(c)}$ which determines the portion of neuron $\mathbf{r}$ assigned to the respective classes. During training, prototype locations $\mathbf{w}_\mathbf{r}$ and label vectors $\mathbf{x}_i$ are adapted according to the given labeled training data. For this purpose, we extend the cost function of the SOM as defined in (5) to a cost function for fuzzy-labeled SOM (FLSOM) by

$$E_{\mathrm{FLSOM}} = (1-\beta)\, E_{\mathrm{SOM}} + \beta E_{\mathrm{FL}} \tag{6}$$

where the factor $\beta \in [0,1]$ 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. As above, E_{SOM} measures the quantization of the map taking topological constraints into account. E_{FL} measures the error of the classification. We choose

$$E_{\mathrm{FL}} = \frac{1}{2} \int P(\mathbf{v}) \sum_{\mathbf{r}} g_\gamma(\mathbf{v}, \mathbf{w}_\mathbf{r}) \left(\mathbf{x} - \mathbf{y}_\mathbf{r}\right)^2 d\mathbf{v} \tag{7}$$

where $g_\gamma(\mathbf{v}, \mathbf{w}_\mathbf{r})$ is a Gaussian kernel describing a neighborhood range in the data space:

$$g_\gamma(\mathbf{v}, \mathbf{w}_\mathbf{r}) = \exp\left(-\frac{\xi(\mathbf{v}, \mathbf{w}_\mathbf{r})}{2\gamma^2}\right). \tag{8}$$

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 $g_\gamma(\mathbf{v}, \mathbf{w}_\mathbf{r})$ depends on the prototype locations, such that E_{FL} is influenced by both $\mathbf{w}_\mathbf{r}$ and $\mathbf{y}_\mathbf{r}$, and an adaptation yields to a different location of prototypes which is also influenced by the labels.

We obtain the update rules by taking the derivatives: Labels are only influenced by the second part E_{FL}, which yields

$$\frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{y_r}} = -\int P(\mathbf{v})\, g_\gamma(\mathbf{v}, \mathbf{w_r})(\mathbf{x} - \mathbf{y_r})\, d\mathbf{v} \tag{9}$$

and the corresponding learning rule

$$\triangle \mathbf{y_r} = \epsilon_l \beta \cdot g_\gamma(\mathbf{v}, \mathbf{w_r})(\mathbf{x} - \mathbf{y_r}) \tag{10}$$

with learning rate $\epsilon_l > 0$. This yields to a weighted average of the data fuzzy labels of those data close to the associated prototypes. However, in comparison to the usual SOM the receptive fields are different because the prototype update is determined by the gradient of (6) which yields $\frac{\partial E_{\mathrm{SOM}}}{\partial \mathbf{w_r}} + \frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{w_r}}$ where

$$\frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{w_r}} = -\frac{1}{4\gamma^2} \int P(\mathbf{v})\, g_\gamma(\mathbf{v}, \mathbf{w_r}) \frac{\partial \xi(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}} (\mathbf{x} - \mathbf{y_r})^2\, d\mathbf{v} \tag{11}$$

which takes the accuracy of fuzzy labeling into account for the weight update. The update rule for the weights thus becomes

$$\begin{aligned} \triangle\, \mathbf{w_r} = &-\epsilon(1-\beta) \cdot h_\sigma(\mathbf{r}, s(\mathbf{v})) \frac{\partial \xi(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}} \\ &+ \epsilon\beta \frac{1}{4\gamma^2} \cdot g_\gamma(\mathbf{v}, \mathbf{w_r}) \frac{\partial \xi(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}} (\mathbf{x} - \mathbf{y_r})^2 . \end{aligned} \tag{12}$$

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 probabilistic class labels between neighbored grid positions.

4 Relevance Learning

As mentioned above, $\xi(\mathbf{v}, \mathbf{w_r})$ is often chosen as squared Euclidean metric such that the term $\frac{\partial \xi(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}}$ becomes $-2(\mathbf{v} - \mathbf{w_r})$. However, the integration of adaptive relevance factors (metric parameters) seems particularly interesting because of an increased flexibility and interpretability of the model with almost the same cost as for the standard metric [5]. Generally, we consider a parametrized 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 relevance learning is to optimize the relevance factors $\boldsymbol{\lambda}$ of the distance measure with respect to the classification task [4],[3], i.e. we consider $\frac{\partial E_{\mathrm{FLSOM}}}{\partial \lambda_l}$. Formal derivation yields

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

with

$$\frac{\partial E_{\text{SOM}}}{\partial \lambda_l} = \frac{1}{2} \int P(\mathbf{v}) \sum_{\mathbf{r}} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}'} h_\sigma(\mathbf{r}, \mathbf{r}') \cdot \frac{\partial \xi^\lambda(\mathbf{v}, \mathbf{w}_\mathbf{r})}{\partial \lambda_l} d\mathbf{v} \tag{14}$$

and

$$\frac{\partial E_{\text{FL}}}{\partial \lambda_l} = -\frac{1}{4\gamma^2} \int P(\mathbf{v}) \sum_{\mathbf{r}} g_\gamma(\mathbf{v}, \mathbf{w}_\mathbf{r}) \frac{\partial \xi^\lambda(\mathbf{v}, \mathbf{w}_\mathbf{r})}{\partial \lambda_l} (\mathbf{x} - \mathbf{y}_\mathbf{r})^2 d\mathbf{v} \tag{15}$$

for the respective parameter adaptation.

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

$$\xi^\lambda(\mathbf{v}, \mathbf{w}) = \sum_i \lambda_i (v_i - w_i)^2 \tag{16}$$

(with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$), relevance learning ranks the input dimensions i according to their relevance for the classification task at hand. Thus, $\frac{\partial \xi(\mathbf{v}, \mathbf{w}_i)}{\partial \mathbf{w}_i}$ becomes

$$\frac{\partial \xi(\mathbf{v}, \mathbf{w}_i)}{\partial \mathbf{w}_i} = -2 \cdot \Lambda \cdot (\mathbf{v} - \mathbf{w}_i) \tag{17}$$

with diagonal matrix Λ with i-th diagonal entry λ_i. The corresponding learning rule for the relevance parameters becomes

$$\triangle \lambda_l = -\epsilon_\lambda \frac{1-\beta}{2} \sum_{\mathbf{r}} h_\sigma(s(\mathbf{v}), \mathbf{r}) \cdot (v_l - (w_\mathbf{r})_l)^2 \tag{18}$$

$$+\epsilon_\lambda \frac{\beta}{4\gamma^2} \sum_{\mathbf{r}} g_\gamma(\mathbf{v}, \mathbf{w}_\mathbf{r})(v_l - (w_\mathbf{r})_l)^2 (\mathbf{x} - \mathbf{y}_\mathbf{r})^2 \tag{19}$$

(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 $\sum_i \lambda_i = 1$.

5 Experiments

5.1 Data Set

In order to demonstrate the practical properties of the proposed algorithm, a quite challenging application in the field of biological image segmentation has been chosen. Against the background of spatiotemporal 3-D modelling of cereal seeds based on up to 2.000 high-resolution images of histological cross-sections a processing transition from crisp to fuzzy segmentation is desirable.

For this purpose all image pixels are characterized by an extensive feature vector containing information on color, geometry and symmetry (such as Cartesian and polar coordinates, distance to centroid, absolute angle to symmetry axis) and particularly texture according to varying neighborhoods (such as Gaussian filters, histogram based features) and subsequently sorted into classes by a suitable classifier. A number of training pixels are used to set up a classification system. These training pixels commonly have a unique class label indicating that tissue which has previously been assigned by an expert. Since this assignment is often not univocal, due to slight transitions from one biological tissue to an adjacent one, some fuzzy image segmentation would retain much more biological knowledge. Further details about the biological background along with fuzzy segmentation using Fuzzy Labelled Neural Gas (FLNG) can be found in [2]. Tab. 1 summarizes those details of the data set relevant to this paper.

Table 1. Details of the utilized demonstration data set. The classes are non-uniformly distributed and usually multimodal.

Number of inputs	Number of classes	SOM-grid size	Number of training examples	Number of test examples
170	5	7×7	ca. 10.000	$>$ 10.000.000

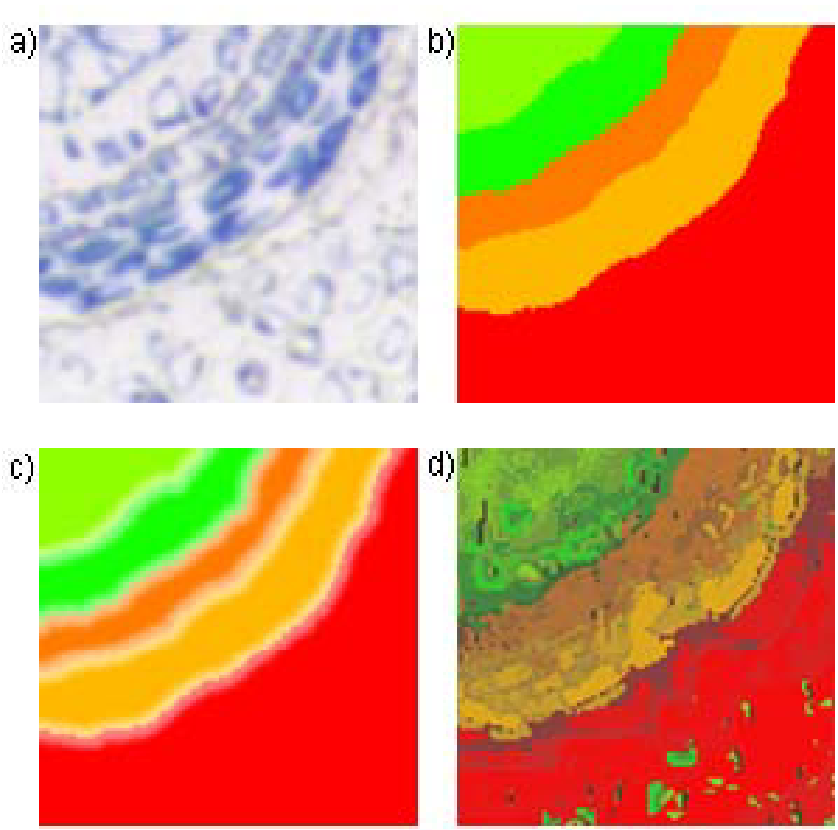

Fig. 2. Corresponding cutouts of images of the same cross-section illustrating the results of an automatic fuzzy classification: a) original colour image, b) manually crisply segmented image, c) manually fuzzily segmented image (see [2]), d) automatic classification using FLSOM.

5.2 Results

Since different areas of the images require more or less fuzzy segmentation, Fig. 2 shows a typical cutout from one complete microscope image containing several transitions of different tissues. In comparison to the crisp segmented image it can clearly be seen, that there are areas with predominantly crisp segmentation as well as areas with mainly fuzzy segmentation. Insofar, the classification system has to reach two partly contradictory goals, the mapping of structural image data onto the classes in a fuzzy manner and the observance of the statistical information about the data distribution. We used the scaled Euclidean distance (16) and a 7×7 SOM grid in the applications, which is chosen according to an optimal topology preserving mapping (the topographic product is approximately zero, indicating good topographic mapping [1]). The learning rate was $\epsilon = 0.01$, $\epsilon_{\boldsymbol{\lambda}} = 0.1\epsilon$ and the balancing parameter $\beta = 0.6$ based on experimental experiences [2],[13]. Relevance learning was incorporated for optimal metric adaptation using the scaled Euclidean metric (16).

The resulted segmentation image based on FLSOM classification is depicted in Fig. 2d. The result shows that the obtained image mixes the original class information overlaid by the structural information (geometry, symmetries ...) contained in the original color image Fig. 2a. This impression is emphasized if the original color image is manually overlaid by the fuzzy classification target (fuzzy labeled) image Fig. 2c and after this compared to the FLSOM classification result. This comparison is shown in Fig. 3, which demonstrates a nice agreement. The segmentation result is comparable to a segmentation obtained by fuzzy labeled neural gas algoritm (FLNG), which also is a neural map based fuzzy classification scheme similar to FLSOM [2],[13]. Favored features of SOMs are the visualization abilities which are also available for FLSOM in advance compared to FLNG. Here, this property is used for investigation of the class structures. For

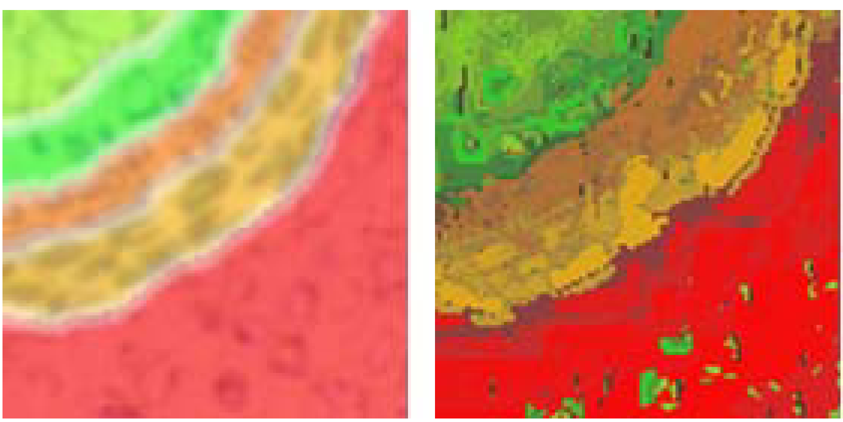

Fig. 3. Comparison of manually overlaid color image Fig. 2a and Fig. 2c (left) with FLSOM resulted classification Fig. 2d (right)

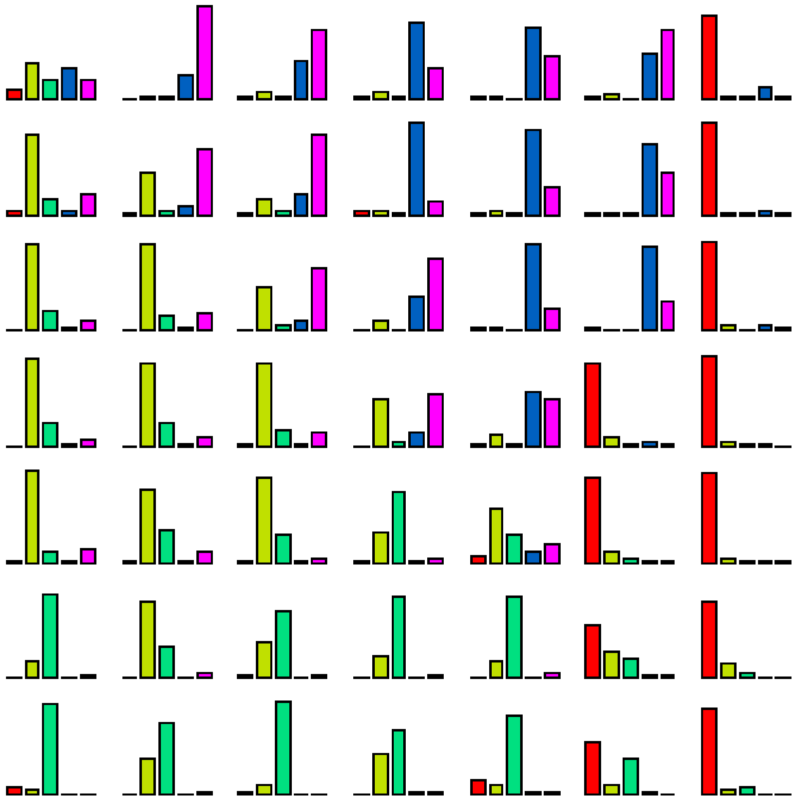

Fig. 4. Distribution of the fuzzy labels of the prototypes according to the grid locations. The topological order of labels performing label clusters can easily be detected.

this purpose, the fuzzy labels of prototypes are plotted according to the underlying topological structure of the 7×7-SOM-grid, Fig. 4. Obviously clear clusters of labels can be locally detected and separated with a smooth change between them. Because of the topology preservation of the SOM-mapping (proven by the topographic product) we can conclude here that a continuos change in the data space leads to a continuos change in the label space and hence classification decision.

Additionally, the adapted relevance profile of the scale parameters λ_i of the scaled Euclidean metric (16) may offer new insights for further investigations, see Fig. 5.This particularly concerns the optimization of the feature vector currently used for the segmentation and the network training, respectively. From the relevance profile it can be concluded that a rather long feature vector is necessary to keep all the information required to distinguish between the different tissues (classes). Nevertheless, the used feature vector is subject to a further optimization based on the obtained relevance profiles at different runs. However, this is not trivial, since several runs at similar classification performance may yield

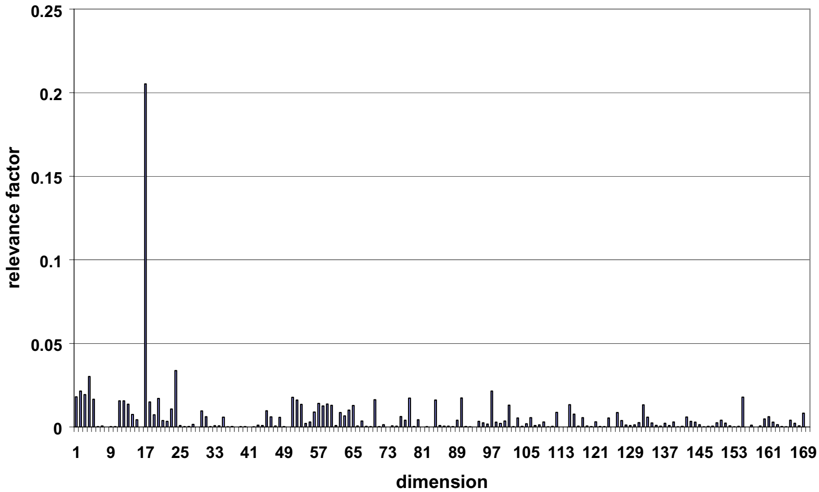

Fig. 5. Relevance profile of the values λ_i for the scaled Euclidean metric used in the application

different relevance profiles. Generally though, this way FLSOM additionally offers a native and self-contained way to keep itself slim.

6 Conclusions

We presented an extension of the SOM for supervised classification tasks, which explicitly adapts the prototypes according to the classification task. It is derived as a gradient descent of a cost function obtained from the SOM formulation according to the Heskes-approach extended by an additional (balanced) term assessing the classification ability. In this way the statistical as well as label properties of the data influence prototype positions and fuzzy label learning. The visualization abilities of SOMs based on the topology preservation property of unsupervised SOMs then can be used for visual inspection of the class labels of the prototypes which may allow a better understanding of the underlying classification decision scheme.

In future work the connections to the unsupervised clustering in SOMs with auxiliary data information should be considered [11]. For this purpose one could interprete the fuzzy class labels as the auxiliary data space.

Acknowledgements

The authors would like to thank Winfriede Weschke for providing expertise on the biological background of the demonstration application. This work was supported by a grant of the German Federal Ministry of Education and Research (No. 0312706A).

References

1. H.-U. Bauer and K. R. Pawelzik. Quantifying the neighborhood preservation of Self-Organizing Feature Maps. *IEEE Trans. on Neural Networks*, 3(4):570–579, 1992.
2. C. Brüß, F. Bollenbeck, F.-M. Schleif, W. Weschke, T. Villmann, and U. Seiffert. Fuzzy image segmentation with fuzzy labeled neural gas. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2006)*, pages 563–568, Brussels, Belgium, 2006. d-side publications.
3. B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.
4. B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.
5. 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.
6. T. Hastie and W. Stuetzle. Principal curves. *J. Am. Stat. Assn.*, 84:502–516, 1989.
7. R. Hecht-Nielsen. Counterprogagation networks. *Appl. Opt.*, 26(23):4979–4984, December 1987.
8. R. Hecht-Nielsen. Applications of counterpropagation networks. *Neural Networks*, 1(2):131–139, 1988.
9. T. Heskes. Energy functions for self-organizing maps. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 303–316. Elsevier, Amsterdam, 1999.
10. T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).
11. J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.
12. 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.
13. T. Villmann, B. Hammer, F.-M. Schleif, and T. Geweniger. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, page in press, 2006.