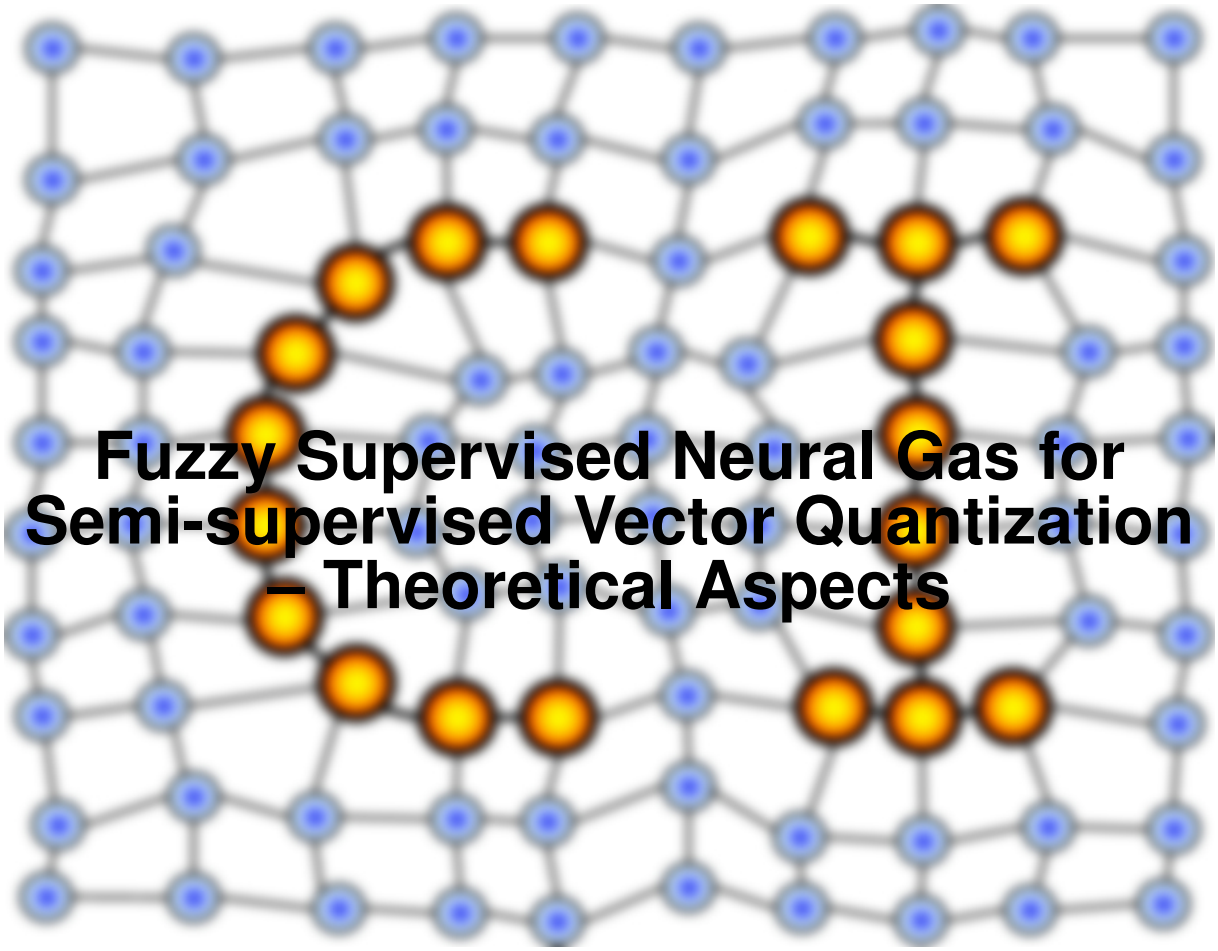
MACHINE LEARNING REPORTS



Fuzzy Supervised Neural Gas for Semi-supervised Vector Quantization – Theoretical Aspects

Report 02/2011 - 2nd extended and revised version
Submitted: 04.05.2011, Revised: 25.08.2011,
Published:29.08.2011

Marika Kästner[1] and Thomas Villmann[1]
(1) University of Applied Sciences Mittweida,
Faculty of Mathematics, Natural and Computer Science,
Computational Intelligence Group, Technikumplatz 17, 09648 Mittweida, Germany
{kaestner,villmann@hs-mittweida.de}

Abstract

In this paper we propose a new approach to combine unsupervised and supervised vector quantization for clustering and fuzzy classification using the framework of neural gas vector quantizer. For this purpose the original cost function is modified in such a way that both aspects, vector quantization and classification, are incorporated. The theoretical justification of the convergence of the new algorithm is given by an adequate redefinition of the underlying dissimilarity measure, which allows a gradient descent learning as known for the original neural gas algorithm. Thus a semi-supervised learning scheme is obtained, which can be interpreted as an association learning. This idea can also be applied for semi-supervised learning of self-organizing maps.

Fuzzy Supervised Neural Gas for Semi-supervised Vector Quantization – Theoretical Aspects

– 2nd extended and revised version –

Marika Kästner and Thomas Villmann

University of Applied Sciences Mittweida,
Faculty of Mathematics, Natural and Computer Sciences,
Computational Intelligence Group,
Technikumplatz 17, 09648 Mittweida, Germany,

email: {kaestner,villmann}@hs-mitttweida.de

Abstract

In this paper we propose a new approach to combine unsupervised and supervised vector quantization for clustering and fuzzy classification using the framework of the neural gas vector quantizer. For this purpose the original cost function is modified in such a way that both aspects, vector quantization and classification, are incorporated. The theoretical justification of the convergence of the new algorithm is given by an adequate redefinition of the underlying dissimilarity measure, which allows a gradient descent learning as known for the original neural gas algorithm. Thus a semi-supervised learning scheme is obtained, which can be interpreted as an association learning. This idea can also be applied for semi-supervised learning of self-organizing maps.

1 Introduction

Unsupervised and supervised vector quantization by neural maps is still an important issue. Neural maps are prototype based algorithms inspired by biological neural systems. Prominent models are the self-organizing map (SOM) and the neural gas network (NG) [5],[7]. These approaches are designed for data clustering (NG) and visualization (SOM). Supervised learning vector quantization follows the idea of prototype based classification preserving the concept of data typical representation in contrast to support vector machines, which emphasize the class borders to describe data classes. Well known such models are the family of learning vector quantizers (LVQ) based on a heuristic adaptation scheme [5], or their cost function based counterpart named generalized LVQ (GLVQ) [10].

There exist only a few approaches to combine unsupervised and supervised learning in SOM or NG. The most intuitive one is a simple post-labeling after unsupervised training. An approach based on a modification of the cost function of NG and SOM (in the HESKES variant, [3]) are the Fuzzy labeled NG (FLNG) and the Fuzzy Labeled SOM (FLSOM) [14, 16]. Both approaches add an extra term to the standard cost function judging the classification accuracy of the prototypes, which are equipped with a class label to be adapted during the learning together with the prototype positions. Yet, the theoretical justification is tricky.

In this paper we propose a much simpler ansatz: We incorporate the classification error in the standard cost functions of NG by a multiplicative factor. Thereby, this factor evaluates the classification accuracy based on a quasi metric [9]. This allows a redefinition of the data metric in such a way that the problem can be handled in this new quasi-metric space analogously to the original NG equipped with the Euclidean metric. Thus the structural framework of standard neural gas is preserved and its convergence properties are transferred to the new model. The new approach can be seen as a kind of association learning known from [8].

This idea can be analogously transferred to the SOM model using the Heskes variant [3]. Overall, the new approach can be seen as a kind of semi-supervised learning. Moreover, the model can be applied to both crisp and fuzzy labeled data.

2 The Fuzzy Supervised Neural Gas Model

The usual neural gas model assumes data points $\mathbf{v} \in V \subset \mathbb{R}^n$ with the data density $P(\mathbf{v})$ and prototypes $\mathbf{w}_j \in \mathbb{R}^n$, $j = 1 \ldots N$. The cost function to be minimized in

NG is

$$E_{\mathrm{NG}} = \sum_j \int P(\mathbf{v})\, h_\sigma^{NG}\left(k_j(\mathbf{v}, \mathbf{w}_j)\right) d(\mathbf{v}, \mathbf{w}_j)\, d\mathbf{v} \tag{1}$$

with a differentiable (in the second argument) dissimilarity measure $d(\mathbf{v}, \mathbf{w}_j)$ usually taken as the Euclidean distance [7]. The function

$$h_\sigma^{NG}\left(k_j(\mathbf{v}, \mathbf{w}_j)\right) = \exp\left(-\frac{k_j(\mathbf{v}, \mathbf{w}_j)}{2\sigma^2}\right) \tag{2}$$

is the neighborhood function depending on the winner rank

$$k_j(\mathbf{v}, \mathbf{w}_j) = \sum_{i=1}^{N} \Theta\left(d(\mathbf{v}, \mathbf{w}_j) - d(\mathbf{v}, \mathbf{w}_i)\right) \tag{3}$$

where

$$\Theta(x) = \begin{cases} 0 & if \quad x \leq 0 \\ 1 & else \end{cases} \tag{4}$$

is the Heaviside function. We remark that $h_\sigma^{NG}\left(k_j(\mathbf{v}, \mathbf{w}_j)\right)$ is evaluated in the data space V. An input vector $\mathbf{v}$ is mapped onto a prototype s by the winner-take-all mapping rule

$$s = \mathrm{argmin}_j\left(d(\mathbf{v}, \mathbf{w}_j)\right) \tag{5}$$

and the learning takes place as stochastic gradient descent $\frac{\partial_S E_{\mathrm{NG}}}{\partial \mathbf{w}_j}$ on E_{NG} according to

$$\triangle \mathbf{w}_j = -h_\sigma^{NG}\left(k_j(\mathbf{v}, \mathbf{w}_j)\right) \cdot \frac{\partial d(\mathbf{v}, \mathbf{w}_j)}{\partial \mathbf{w}_j}. \tag{6}$$

In the following we develop a new variant of standard NG, which integrates additional class information into the standard model ending up with a (semi-) supervised variant of standard NG, which is also applicable to fuzzy classification problems and therefore denoted as Fuzzy Supervised Neural Gas – (FSNG).

2.1 The FSNG-Model

First in this section, we shortly mention the earlier approach to deal with fuzzy labeled data learning in NG - the Fuzzy Labeled Neural Gas (FLNG) and point out its difficulties. Second, we turn to the new FSNG model, which overcome some of these problems.

2.1.1 Earlier approaches - the Fuzzy Labeled Neural Gas – FLNG

We start with a brief introduction of FLNG as it was introduced in [14]. We suppose C data classes. Each data vector $\mathbf{v}$ is accompanied by a data assignment vectors $\mathbf{c_v} \in [0,1]^C$ with vector entries taken as class probability or possibility assignments. Analogously, we also equip the prototypes $\mathbf{w}_j$ with class labels $\mathbf{y}_j$. The original cost function of NG (1) is extended in FLNG by an additional term judging the classification ability:

$$E_{\mathrm{FLNG}} = (1-\gamma)\, E_{\mathrm{NG}} + \gamma E_{\mathrm{FL}} \tag{7}$$

where

$$E_{\mathrm{FL}} = \sum_j \int P(\mathbf{v})\, g(\mathbf{v}, \mathbf{w}_j)\, \delta(\mathbf{c_v}, \mathbf{y}_j)\, d\mathbf{v}$$

with a dissimilarity measure $\delta(\mathbf{c_v}, \mathbf{y}_j)$ for the class assignment vectors. The parameter $\gamma \in [0,1]$ determines the influence of the class information with $\gamma = 0$ yields the standard NG. Hence, the cost function E_{FLNG} can be rewritten as

$$E_{\mathrm{FLNG}} = \sum_j \int P(\mathbf{v}) \left[h_\sigma^{NG}(k_j(\mathbf{v}, \mathbf{w}_j))\,(1-\gamma)\, d(\mathbf{v}, \mathbf{w}_j) + \gamma g(\mathbf{v}, \mathbf{w}_j)\, \delta(\mathbf{c_v}, \mathbf{y}_j) \right] d\mathbf{v} \tag{8}$$

The neighborhood cooperativeness function $g(\mathbf{v}, \mathbf{w}_j)$ for the label accuracy in FLNG explicitly takes into account the dissimilarity $d(\mathbf{v}, \mathbf{w}_j)$ between the prototype $\mathbf{w}_j$ and the data vector $\mathbf{v}$ but has to be defined differently for discrete and continuous data distributions. For continuous data the neighborhood function becomes

$$g_{cont}(\mathbf{v}, \mathbf{w}_j) = \exp\left(-\frac{d(\mathbf{v}, \mathbf{w}_j)}{2\sigma_g^2} \right)$$

whereas

$$g_{discr}(\mathbf{v}, \mathbf{w}_j) = h_\sigma^{NG}(k_j(\mathbf{v}, \mathbf{w}_j))$$

is valid for the discrete setting. In the latter case, the cost function (8) can be further simplified to

$$E_{\mathrm{FLNG}} = \sum_j \int P(\mathbf{v})\, h_\sigma^{NG}(k_j(\mathbf{v}, \mathbf{w}_j))\, \widetilde{D}(\mathbf{v}, \mathbf{w}_j)\, d\mathbf{v}$$

with a new *additive* distortion measure

$$\widetilde{D}(\mathbf{v}, \mathbf{w}_j) = \left[(1-\gamma)\, d(\mathbf{v}, \mathbf{w}_j) + \gamma \delta(\mathbf{c_v}, \mathbf{y}_j) \right]. \tag{9}$$

The necessary for this distinction in the neighborhood cooperativeness for the labels is a consequence to assure a valid convergence proof of the algorithm, for details we refer to [14]. We will see in the following that the new FSNG proposed here is not affected by such difficulties.

2.1.2 The new Fuzzy Supervised Neural Gas for class association learning

For the FSNG model, we now consider the cost function

$$E_{\mathrm{FSNG}} = \sum_j \int P(\mathbf{v})\, h_\sigma^{NG}\left(k_j(\mathbf{v}, \mathbf{w}_j)\right) D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma)\, d\mathbf{v} \tag{10}$$

which is structurally similar to standard neural gas. As for the additive dissimilarity $\widetilde{D}(\mathbf{v}, \mathbf{w}_j)$, the new deviation measure $D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)$, describing the dissimilarity between data and prototype vectors, takes into account both the usual dissimilarity $d(\mathbf{v}, \mathbf{w}_j)$ between data and prototypes as well as their dissimilarity $\delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j)$ for the class information as introduced for FLNG (9). In the simplest case, both measures, $d(\mathbf{v}, \mathbf{w}_j)$ and $\delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j)$, could be chosen as the Euclidean distance. In distinction to FLNG, we propose for the FSNG a *multiplicative* combination

$$D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma) = (\gamma \cdot \delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j) + \varepsilon_\delta) \cdot ((1-\gamma) \cdot d(\mathbf{v}, \mathbf{w}_j) + \varepsilon_d) - \varepsilon_\delta \varepsilon_d \tag{11}$$

of both dissimilarity measures. Again, the parameter $\gamma \in [0, 1]$ determines the influence of the class information with $\gamma = 0$ yielding the standard NG. The additional parameter vector $\varepsilon = (\varepsilon_\delta, \varepsilon_d)$ is necessary in D_ε to prevent unexpected behavior of the FSNG under certain conditions, which are discussed more detailed later.

Yet, the winner determination rule (5) now becomes

$$s = \mathrm{argmin}_j \left(D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma)\right) \tag{12}$$

in this FSNG model during the learning, with the winner rank function now rewritten as

$$k_j^\gamma(\mathbf{v}, \mathbf{w}_j) = \sum_{i=1}^{N} \Theta\left(D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma) - D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)\right) \tag{13}$$

compared to that (3) of the original NG. In the recall phase, when classification is carried out and, hence, no label information is available, the standard winner rule (5) of NG is applied.

As in FLNG, the FSNG model leads to a prototype adaptation influenced by the class agreement $\delta(\mathbf{c_v}, \mathbf{y}_j)$:

$$\triangle \mathbf{w}_j = -(1-\gamma)\left(\gamma \cdot \delta(\mathbf{c_v}, \mathbf{y}_j) + \varepsilon_\delta\right) \cdot h_\sigma^{NG}\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right) \cdot \frac{\partial d(\mathbf{v}, \mathbf{w}_j)}{\partial \mathbf{w}_j} \tag{14}$$

accompanied by a label adaptation

$$\triangle \mathbf{y}_j = -\gamma \cdot \left((1-\gamma) \cdot d(\mathbf{v}, \mathbf{w}_j) + \varepsilon_d\right) \cdot h_\sigma^{NG}\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right) \cdot \frac{\partial \delta(\mathbf{c_v}, \mathbf{y}_j)}{\partial \mathbf{y}_j} \tag{15}$$

such that both, prototype vectors and their class assignment vectors, are parallely adapted.

The merging of data and class dissimilarity into a single dissimilarity measure for learning classification was first proposed in the model Learning of Associations by Self-Organization (LASSO,[8]). In this approach, originally introduced for SOM but analogously applicable to NG, modified data vectors $\hat{\mathbf{v}} = (\mathbf{v} \oplus \mathbf{c_v}) \in \widehat{V} \subseteq \mathbb{R}^n \times \mathbb{R}^C$ are generated with $\oplus$ being the concatenation operation. The prototypes $\widehat{\mathbf{w}}_j \in \mathbb{R}^n \times \mathbb{R}^C$ in this model have the same structure. Learning the associations in the LASSO model takes place as usual SOM learning using the Euclidean distance but now between the data $\hat{\mathbf{v}}$ and the prototypes $\widehat{\mathbf{w}}_j$. In the recall phase however, when no label information is available, the Euclidean distance is calculated only with respect to the original data vectors $\mathbf{v}$ as in FSNG. Yet, the FSNG approach offer a greater flexibility for association lerning due to the possibility of appropriate balancing of unsupervised and supervised information by means of the balancing parameter γ. Further, the parameter vector $\varepsilon = (\varepsilon_\delta, \varepsilon_d)$ plays an essential role in case of a perfect match for prototype learning but remaining insufficient classification accuracy and vice versa, as explained in detail in the next section.

2.2 Properties of the dissimilarity measure $D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)$

We will show later in this paper that FSNG adaptation performs a stochastic gradient descent learning for the FSNG cost function E_{FSNG} (10). This stochastic gradient descent with respect to the prototypes $\mathbf{w}_i$ and their class label vectors $\mathbf{y}_i$ takes place for a given data vector $\mathbf{v}$ and its class assignment $\mathbf{c_v}$ proportionally to the partial derivatives

$$\frac{\partial_S E_{\mathrm{FSNG}}}{\partial \mathbf{w}_i} = \frac{\partial_S E_{\mathrm{FSNG}}}{\partial D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)} \frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)}{\partial \mathbf{w}_i} \tag{16}$$

and

$$\frac{\partial_S E_{\mathrm{FSNG}}}{\partial \mathbf{y}_i} = \frac{\partial_S E_{\mathrm{FSNG}}}{\partial D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)} \frac{\partial D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)}{\partial \mathbf{y}_i} \tag{17}$$

Therfore, we investigate the partial derivatives of $D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)$ with respect to $\mathbf{w}_i$ and $\mathbf{y}_i$ in more detail:

$$\frac{\partial D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)}{\partial \mathbf{w}_i} = (1-\gamma) \cdot \left(\gamma \cdot \delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_i\right) + \varepsilon_\delta\right) \cdot \frac{\partial d\left(\mathbf{v}, \mathbf{w}_i\right)}{\partial \mathbf{w}_i} \tag{18}$$

and

$$\frac{\partial D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)}{\partial \mathbf{y}_i} = \gamma \cdot \left((1-\gamma) \cdot d\left(\mathbf{v}, \mathbf{w}_i\right) + \varepsilon_d\right) \cdot \frac{\partial \delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_i\right)}{\partial \mathbf{y}_i} \tag{19}$$

determining the update formula (14) and (15). If the quadratic Euclidean distance is used for $d\left(\mathbf{v}, \mathbf{w}_j\right)$ and $\delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_j\right)$, we immediately find

$$\frac{\partial d\left(\mathbf{v}, \mathbf{w}_i\right)}{\partial \mathbf{w}_i} = -2\left(\mathbf{v} - \mathbf{w}_i\right) \tag{20}$$

and

$$\frac{\partial \delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_i\right)}{\partial \mathbf{y}_i} = -2\left(\mathbf{c}_\mathbf{v} - \mathbf{y}_i\right) \tag{21}$$

for prototype and class assignment adaptation, respectively.

It should be mentioned that $D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)$ is not a standard (mathematical) metric since it violates the triangle inequality. However, $D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)$ fulfills the requirements of a quasi-metric [9]. In particular, we have $D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right) = 0$ for a perfect match of the prototype as well as its label.

For learning in FSNG we have to distinguish the following extreme cases, which should be of special interest:

1. $d\left(\mathbf{v}, \mathbf{w}_i\right) = 0$ and $\delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_i\right) \neq 0$, i.e. the prototype is perfectly placed but its label is not adequate: In that case a non-vanishing term

$$\frac{\partial D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)}{\partial \mathbf{y}_i}|_{d(\mathbf{v}, \mathbf{w}_i)=0} = \gamma \cdot \varepsilon_d \cdot \frac{\partial \delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_i\right)}{\partial \mathbf{y}_i} \tag{22}$$

   remains, which guarantees the label adaptation.

2. $d\left(\mathbf{v}, \mathbf{w}_i\right) \neq 0$ and $\delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_i\right) = 0$, i.e. the prototype label perfectly matches but the prototype itself is not optimally adjusted: In that case

$$\frac{\partial D_\varepsilon\left(\mathbf{v}, \mathbf{w}_i, \gamma\right)}{\partial \mathbf{w}_i}|_{\delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_i)=0} = (1-\gamma) \cdot \varepsilon_\delta \cdot \frac{\partial d\left(\mathbf{v}, \mathbf{w}_i\right)}{\partial \mathbf{w}_i} \tag{23}$$

   is non-vanishing such that prototype learning is still possible.

2.3 The FSNG algorithm as a stochastic gradient

We show in this section that the adaptation dynamic of prototypes (14) and labels (15) of FSNG follows a stochastic gradient descent on the cost function given in (10). Thus it overcomes the difficulties in the convergence proof of FLNG, where we have to distinguish discrete and continuous data distributions [14].

Following the original work of MARTINETZ ET AL. [7] we have to investigate for convergence of FSNG the derivatives $\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{w}_i}$ and $\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{y}_i}$.

We start with consideration of the prototype dynamic. We have

$$\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{w}_i} = R_i + \int P(\mathbf{v})\, h_\sigma^{NG}\left(k_i^\gamma(\mathbf{v}, \mathbf{w}_i)\right) \frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)}{\partial \mathbf{w}_i} d\mathbf{v} \tag{24}$$

and the derivative $\frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)}{\partial \mathbf{w}_i}$ is taken from (18). The term R_i is obtained as

$$R_i = \sum_j \int P(\mathbf{v}) \frac{\partial h_\sigma^{NG}\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right)}{\partial \mathbf{w}_i} D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma)\, d\mathbf{v} \tag{25}$$

with

$$\frac{\partial h_\sigma^{NG}\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right)}{\partial \mathbf{w}_i} = \left[h_\sigma^{NG}\right]'\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right) \cdot \frac{\partial k_j^\gamma(\mathbf{v}, \mathbf{w}_j)}{\partial \mathbf{w}_i}$$

and $\left[h_\sigma^{NG}\right]'(\bullet)$ denotes the derivative of $h_\sigma^{NG}(\bullet)$. If R_i is vanishing, then the derivative (24) yields the prototype learning rule (14) of FSNG. We decompose R_i into $R_i = R_{i,1} + R_{i,2}$ such that

$$R_{i,1} = \int P(\mathbf{v}) \left[h_\sigma^{NG}\right]'\left(k_i^\gamma(\mathbf{v}, \mathbf{w}_i)\right) \cdot D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma) \frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)}{\partial \mathbf{w}_i} \sum_l \theta(\triangle_{il})\, d\mathbf{v}$$

and

$$-R_{i,2} = \sum_j \int P(\mathbf{v}) \left[h_\sigma^{NG}\right]'\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right) \cdot D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma) \cdot \frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)}{\partial \mathbf{w}_i} \cdot \theta(\triangle_{ji})\, d\mathbf{v}$$

with $\triangle_{mk} = D_\varepsilon(\mathbf{v}, \mathbf{w}_m, \gamma) - D_\varepsilon(\mathbf{v}, \mathbf{w}_k, \gamma)$. Thereby we have used the fact that the derivative of the Heaviside function $\Theta(x)$ from (4) is the Dirac distribution $\theta(x)$, which is zero iff $x \neq 0$ and $\int \theta(x)\, dx = 1$. We emphasize at this point again that the rank function $k_j^\gamma(\mathbf{v}, \mathbf{w}_j)$ depends on the new dissimilarities $D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma)$.

For $R_{i,2}$ we can interchange integration and summation. Further, $R_{i,2}$ is non-vanishing only for $\triangle_{ij} = 0$ according to the Dirac functional θ, which is equivalent to $D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma) = D_\varepsilon(\mathbf{v}, \mathbf{w}_j, \gamma)$. For those $\mathbf{v}$'s obviously the equation

$$\sum_k \theta(\triangle_{ik}) = \sum_k \theta(\triangle_{jk})$$

holds implying immediately the equivalence $k_j^\gamma(\mathbf{v},\mathbf{w}_j) = k_i^\gamma(\mathbf{v},\mathbf{w}_i)$. At this end, we obtain

$$-R_{i,2} = \int P(\mathbf{v}) \left[h_\sigma^{NG}\right]'(k_i^\gamma(\mathbf{v},\mathbf{w}_i)) \cdot D_\varepsilon(\mathbf{v},\mathbf{w}_i,\gamma) \cdot \frac{\partial D_\varepsilon(\mathbf{v},\mathbf{w}_i,\gamma)}{\partial \mathbf{w}_i} \cdot \sum_l \theta(\triangle_{il})\, d\mathbf{v}$$

which leads to $R_{i,1} = -R_{i,2}$ paying attention to the fact that $\theta(x)$ is symmetric: $\theta(x) = \theta(-x)$.

Further, using the derivative $\frac{\partial D_\varepsilon(\mathbf{v},\mathbf{w}_i,\gamma)}{\partial \mathbf{w}_i}$ from (18), the gradient $\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{w}_i}$ in (24) finally reduces to

$$\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{w}_i} = (1-\gamma) \int P(\mathbf{v})\, h_\sigma^{NG}(k_i(\mathbf{v},\mathbf{w}_i))\, (\gamma \cdot \delta(\mathbf{c}_\mathbf{v},\mathbf{y}_i) + \varepsilon_\delta) \cdot \frac{\partial d(\mathbf{v},\mathbf{w}_i)}{\partial \mathbf{w}_i} d\mathbf{v}, \tag{26}$$

which is exactly the averaged prototype dynamic (14). This completes the proof for the prototype dynamic.

It remains to investigate the dynamic for the class labels $\mathbf{y}_i$. We have

$$\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{y}_i} = \tilde{R}_i + \gamma \int P(\mathbf{v})\, h_\sigma^{NG}(k_i^\gamma(\mathbf{v},\mathbf{w}_i)) \frac{\partial D_\varepsilon(\mathbf{v},\mathbf{w}_i,\gamma)}{\partial \mathbf{y}_i} d\mathbf{v}$$

with

$$\tilde{R}_i = \sum_j \int P(\mathbf{v}) \frac{\partial h_\sigma^{NG}\left(k_j^\gamma(\mathbf{v},\mathbf{w}_j)\right)}{\partial \mathbf{y}_i} D_\varepsilon(\mathbf{v},\mathbf{w}_j,\gamma)\, d\mathbf{v}.$$

In complete analogy we find $\tilde{R}_i$ vanishing, too. Thus, we obtain

$$\frac{\partial E_{\mathrm{FSNG}}}{\partial \mathbf{y}_i} = \gamma \int P(\mathbf{v})\, h_\sigma^{NG}(k_i(\mathbf{v},\mathbf{w}_i))\, ((1-\gamma) \cdot d(\mathbf{v},\mathbf{w}_i) + \varepsilon_d) \cdot \frac{\partial \delta(\mathbf{c}_\mathbf{v},\mathbf{y}_i)}{\partial \mathbf{y}_i} d\mathbf{v}$$

for the averaged label dynamic corresponding to (15). This completes the proof for the desired FSNG dynamic.

2.4 Semi-supervised learning – balancing between unsupervised and supervised learning by the parameter γ

The quasi-metric $D_\varepsilon(\mathbf{v},\mathbf{w}_i,\gamma)$ depends on the balancing parameter γ weighting the unsupervised and supervised aspects. Experiences from earlier models (Fuzzy Label Neural Gas – FLNG, [14]) suggest a careful control of this parameter beginning with $\gamma(0) = 0$ and later (adiabatic) increase up to a final value $\gamma_{\max}$, which should be chosen as $\gamma_{\max} < 1$ to avoid instabilities as known from FLNG. This can be interpreted as a remaining influence of unsupervised learning in the supervised learning phase of FSNG.

3 Fuzzy Supervised SOM

In this section we extend the above semi-supervised learning ideas to the Heskes variant of SOMs, which also performs a gradient descent learning.

For a SOM we assume in the following that the index $\mathbf{r}$ of a prototype $\mathbf{w_r}$ refers to a neuron $\mathbf{r} \in A$, whereby A is equipped with a underlying topological structure usually chosen as a regular grid. We denote the grid distance between nodes $\mathbf{r}$ and $\mathbf{r}'$ by $d_A(\mathbf{r}, \mathbf{r}')$. The original SOM introduced by T. KOHONEN, which is based of the same mapping rule (5) as NG, does not follow a stochastic gradient of a cost function [1]. Yet, a slight modification of this rule allows this desired result [3]:

$$\mathbf{s}(\mathbf{v}) = \text{argmin}_{\mathbf{r} \in A} \left(\sum_{\mathbf{r}' \in A} h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') \, d(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \right) \tag{27}$$

with

$$h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') = \exp\left(-\frac{d_A(\mathbf{r}, \mathbf{r}')}{2\sigma^2} \right) \tag{28}$$

as neighborhood function, but now determined on the neuron grid A. Following the ansatz from *T. Heskes*, we denote

$$e_{\mathbf{r}}(\mathbf{v}) = \sum_{\mathbf{r}' \in A} h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') \, d(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \tag{29}$$

as local costs for neuron $\mathbf{r}$ given the input $\mathbf{v}$ such that (27) can be rewritten as

$$\mathbf{s}(\mathbf{v}) = \text{argmin}_{\mathbf{r} \in A} \left(e_{\mathbf{r}}(\mathbf{v}) \right). \tag{30}$$

Then a cost function for SOM can be defined by

$$E_{\text{SOM}} = \int P(\mathbf{v}) \, e_{\mathbf{s}(\mathbf{v})}(\mathbf{v}) \, d\mathbf{v} \tag{31}$$

which leads to the stochastic gradient learning

$$\triangle \mathbf{w_r} \sim -\sum_{\mathbf{r}' \in A} h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') \frac{\partial d(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}} \tag{32}$$

in complete analogy to the NG. The subtle but essential distinction of the Heskes-SOM compared to the original SOM is the mapping based on local costs, which have to be the basis of the cost function as defined in (31): The derivation of the gradient descent learning is only valid iff the local costs in the cost function (31) are exactly the same as those used for the mapping.

Hence, we can replace the dissimilarity measure $d(\mathbf{v}, \mathbf{w_r})$ by $D_\varepsilon(\mathbf{v}, \mathbf{w}_i, \gamma)$ in the local costs (29) an feed these into the cost function (31). This change we have to apply also to the mapping rule (30) to avoid the violation of necessary condition of gradient descent learning for Heskes-SOMs. Thus we obtain

$$\triangle \mathbf{w_r} \sim -\sum_{\mathbf{r}' \in A} h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') \frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w_r}, \gamma)}{\partial \mathbf{w_r}} \tag{33}$$

and

$$\triangle \mathbf{y_r} \sim -\sum_{\mathbf{r}' \in A} h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}') \frac{\partial D_\varepsilon(\mathbf{v}, \mathbf{w_r}, \gamma)}{\partial \mathbf{y_r}} \tag{34}$$

in complete analogy to FSNG. We refer to this fuzzy supervised SOM as FSSOM.

4 Conclusion

We provide in this paper a new approach for semi-supervised learning in neural gas and self-organizing maps. The new approach combines into one single dissimilarity measure both the dissimilarity between data and prototypes as well as their class dissimilarity in a multiplicative manner. This mixture is balanced using a control parameter γ. We show for NG that the mathematical structure of the underlying cost function is the same than for the original NG, if an adequate redefinition of the dissimilarity measure takes place. In consequence, the theoretical framework of the original algorithm also justifies the new approach. The approach can aslo be transferred to SOMs. However, the theoretical assumptions of stochastic gradient descent learning for an analog SOM modification are only valid for the Heskes variant of SOM.

Obviously, the new approach allows a broad variability of dissimilarity measures d in the data space and δ for the fuzzy labels. Surely, the Euclidean distance is a good choice. However, interesting alternatives are under discussion for different data types at least for the data dissimilarity measures. Prominent examples are the scaled Euclidean metric for relevance learning [2] and their functional counterpart [4], or the Sobolev distance [15] and other functional norms [6], if the data are supposed to be representations of functions. Generalization of the scaled Euclidean metric are quadratic forms used in matrix learning [12]. Divergences are proposed for spectral data as suitable data dissimilarity measures [13], whereas the utilization of differentiable kernel also seems to be a new promising alternative for data dissimilarity judgment [11].

References

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

[2] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

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

[4] M. Kästner and T. Villmann. Functional relevance learning in generalized learning vector quantization. *Machine Learning Reports*, 5(MLR-01-2011):81–89, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_01_2011.pdf.

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

[6] 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.

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

[8] S. Midenet and A. Grumbach. Learning associations by self-organizatiom: the LASSO model. *Neurocomputing*, 6:343–361, 1994.

[9] E. Pekalska and R. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2006.

[10] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[11] F.-M. Schleif, T. Villmann, B. Hammer, P.Schneider, and M. Biehl. Generalized derivative based kernelized learning vector quantization. In C. Fyfe, P. Tino, D. Charles, C. Garcia-Osorio, and H. Yin, editors, *Proceedings of the Conference IDEAL*, volume 6283 of *LNCS*, pages 21–28, Berlin, 2010. Springer.

[12] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.

[13] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

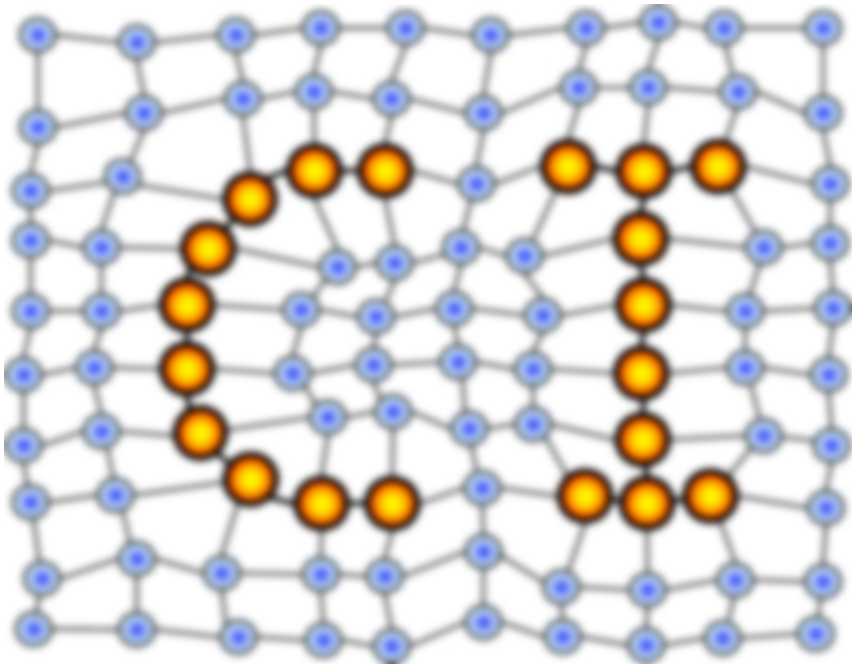
[14] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, and W. Herrmann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19:772–779, 2006.

[15] T. Villmann and F.-M. Schleif. Functional vector quantization by neural maps. In J. Chanussot, editor, *Proceedings of First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2009)*, pages 1–4. IEEE Press, 2009. ISBN 978-1-4244-4948-4.

[16] T. Villmann, F.-M. Schleif, E. Merényi, and B. Hammer. Fuzzy labeled self-organizing maps for classification of spectra. In F. Sandoval, A. Prieto, J. Cabestany, and M. Grana, editors, *Computational and Ambient Intelligence – Proceedings of the 9th Work-conference on Artificial Neural Networks (IWANN), San Sebastian (Spain)*, LNCS 4507, pages 556–563. Springer, Berlin, 2007.

MACHINE LEARNING REPORTS

Report 02/2011 - 2nd extended and revised version



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

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

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

▽ Acknowledgments

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