ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Integration of Structural Expert Knowledge about Classes for Classification Using the Fuzzy Supervised Neural Gas

M. Kästner[1], W. Hermann[2], and T. Villmann[1]

1- University of Appl. Sciences Mittweida - Dept. of Mathematics
Mittweida, Saxonia - Germany

2- Paracelsus-Hospital Zwickau - Dept. of Neurology
Zwickau, Saxonia - Germany

Abstract. In this paper we describe an approach to integrate structural expert knowledge about class relations into classification schemes under the assumption of unary class coding. Exemplary, we show in a medical application such an integration for incorporation of prior medical knowledge about uncertainty for distinguishing patient classes. This knowlegde is integrated here in a class dissimilarity measure used for training the classification model.

1 Introduction

Automatic classification using machine learning tools is a widely accepted methodology. A large variety of classification problems were successfully solved in different application areas ranging from economics, engineering and physics to biology and medicine, to name just a few. Frequently, standard machine learning tools are/can be applied without any or minimum specific knowledge about the structure of the considered problem. This can be seen as one of the key advantages of these tools: their suitability for many different kinds of problems. However, for specific tasks it might be necessary to integrate problem dependent, structural expert knowledge to end up with a model, particularly optimized and specified for a certain classification task. Examples for those specifications could be feature selection, the utilization of more adequate non-standard metrics reflecting particular data properties or integration of uncertainty in data by interval arithmetic of fuzzy approaches. In medicine and psychology, frequently patients are diagnosed by medical doctors based on clinical expert knowledge, which is later not explicitly contained in the labeled data. This knowledge could be additional information about the investigated diseases, the patients, etc., depending on the expert level. Yet, after the medical diagnosis is done, these uncertainties or apriori known disease relations are frequently dropped.

Many machine learning models for classification tasks like multi-layer perceptrons, counter-propagation networks or fuzzy classification schemes assume an unary coding of the class information of the data such that deviations and accuracy information of the desired model output can be expressed as a numerical value. Both problems have in common that apriori known class relations or uncertainties in labeling (diagnosis) would influence the classification decision. Therefore, such information should be used for the classifier training. In this paper, we propose the explicit incorporation of such expert knowledge into machine learning classification systems. In particular, we suggest to code these information into a class similarity/dissimilarity measure, which then is used in a classification model to judge class label agreements. Exemplary we demonstrate this for a semi-supervised vector quantization model - the Fuzzy Supervised

Neural Gas when applying it for classification of patients suffering from a copper metabolism disease based on neurophysiological measurements.

2 The Standard Fuzzy Supervised Neural Gas

As a model for fuzzy classification problems using structural expert knowledge, we extend the Fuzzy Supervised Neural Gas (FSNG,[13]) to deal with this information. The FSNG is a semi-supervised extension of the unsupervised standard neural gas vector quantizer (NG,[8]). It assumes data points $\mathbf{v} \in V \subset \mathbb{R}^n$ with the data density $P(\mathbf{v})$, prototypes $\mathbf{w}_j \in \mathbb{R}^n$, $j = 1 \ldots N$ and a differentiable dissimilarity measure $d(\mathbf{v}, \mathbf{w}_j)$ in the data space. Further, each data vector $\mathbf{v}$ is accompanied by a data assignment vectors $\mathbf{c}_\mathbf{v} \in [0,1]^{N_C}$ with vector entries taken as class probability or possibility assignments. Thus, it can be seen as a variant of unary class coding of the N_C classes. Analogously, we assign to each prototypes $\mathbf{w}_j$ a class label vector $\mathbf{y}_j$. The apriori class structure information as well as expert knowledge about the classes is reflected by a predefined class dissimilarity measure $\delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j)$. The cost function

$$E_{\mathrm{FSNG}} = \frac{1}{K(\sigma)} \sum_j \int P(\mathbf{v})\, h_\sigma\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j, D_\varepsilon)\right) D_\varepsilon(\mathbf{v}, \mathbf{c}_\mathbf{v}, \mathbf{w}_j, \mathbf{y}_j, \gamma)\, d\mathbf{v} \qquad (1)$$

with a differentiable measure $\delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j)$ between the label vectors, has to be minimized in FSNG. It is structurally similar to that of NG but with a new dissimilarity measure

$$D_\varepsilon(\mathbf{v}, \mathbf{c}_\mathbf{v}, \mathbf{w}_j, \mathbf{y}_j, \gamma) = (\gamma \cdot \delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j) + \varepsilon_\delta) \cdot ((1-\gamma) \cdot d(\mathbf{v}, \mathbf{w}_j) + \varepsilon_d) - \varepsilon_\delta \varepsilon_d \quad (2)$$

combining the data and the class dissimilarity in a multiplicative manner. The mixing parameter $\gamma \in [0,1]$ determines the influence of the class information with $\gamma = 0$ yielding the standard NG. The additional parameter vector $\varepsilon = (\varepsilon_\delta, \varepsilon_d)$ prevents zero learning in case of perfect match between the prototype $\mathbf{w}_j$ and data point $\mathbf{v}$ but differing class labels $\mathbf{y}_j$ and $\mathbf{c}_\mathbf{v}$ and vice versa [13]. Further,

$$h_\sigma\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j, D_\varepsilon)\right) = \exp\left(-\frac{k_j^\gamma(\mathbf{v}, \mathbf{w}_j, D_\varepsilon)}{2\sigma^2}\right) \qquad (3)$$

is the neighborhood function of prototypes depending on the dissimilarity differences by the rank function

$$k_j^\gamma(\mathbf{v}, \mathbf{w}_j, D_\varepsilon) = \sum_{i=1}^{N} \Theta\left(D_\varepsilon(\mathbf{v}, \mathbf{c}_\mathbf{v}, \mathbf{w}_j, \mathbf{y}_j, \gamma) - D_\varepsilon(\mathbf{v}, \mathbf{c}_\mathbf{v}, \mathbf{w}_i, \mathbf{y}_i, \gamma)\right) \qquad (4)$$

where the Heaviside function $\Theta(x)$ equals one if $x \leq 0$ and is zero elsewhere. In the FSNG model, both the prototypes as well as their class label vectors are adapted according to a stochastic gradient descent learning. Thereby, the prototype learning is influenced by the class agreement $\delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j)$:

$$\triangle \mathbf{w}_j = -(1-\gamma)(\gamma \cdot \delta(\mathbf{c}_\mathbf{v}, \mathbf{y}_j) + \varepsilon_\delta) \cdot h_\sigma\left(k_j^\gamma(\mathbf{v}, \mathbf{w}_j)\right) \cdot \frac{\partial d(\mathbf{v}, \mathbf{w}_j)}{\partial \mathbf{w}_j} \qquad (5)$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

and the label adaptation

$$\triangle \mathbf{y}_j = -\gamma \cdot \left((1-\gamma) \cdot d\left(\mathbf{v}, \mathbf{w}_j\right) + \varepsilon_d\right) \cdot h_\sigma\left(k_j^\gamma\left(\mathbf{v}, \mathbf{w}_j\right)\right) \cdot \frac{\partial \delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_j\right)}{\partial \mathbf{y}_j} \tag{6}$$

also takes the data dissimilarity $d\left(\mathbf{v}, \mathbf{w}_j\right)$ into account. According to the mixing parameter γ a semi-supervised learning scheme is obtained. In the recall phase, when data class label are not available, the data mapping is realized only on the base of the data dissimilarities $d\left(\mathbf{v}, \mathbf{w}_j\right)$: An input vector $\mathbf{v}$ is mapped onto a prototype s by the winner-take-all mapping rule

$$s = \operatorname{argmin}_j \left(d\left(\mathbf{v}, \mathbf{w}_j\right)\right) \tag{7}$$

and thereafter, as response, the class label associated to that data point simply is the class label vector $\mathbf{y}_s$ of the winning prototype. Due to the lack of space, we refer to [13] for further details.

3 Integration of Expert Knowledge

We now show, how structural expert knowledge can be intergated into the FSNG model. So far, we did not specified the data and class dissimilarities $d\left(\mathbf{v}, \mathbf{w}_j\right)$ and $\delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_j\right)$. Simplest, the Euclidean distance could be applied for both in FSNG. However, in dependence of the shape and structure of the data more adequate dissimilarity measures can be chosen such as the weighted Euclidean distance, divergences or differentiable kernels to name just a few [3, 10, 11, 12]. In this way, prior expert knowledge can be easily fed into the model. If the *data dissimilarity measure* is parametrized and differentiable, then its adaptation can be realized again as stochastic gradient descent as it is known from relevance learning [3], for example.

Obviously, the *class dissimilarity measure* can also be subject of expert knowledge integration. For example, in breast cancer staging mainly five stages of increasing impairment are distinguished (except the healthy stage - without symptoms) : pre-cancerous or non-invasive (level 0), invasive with different sizes, and different state of spreading to axillary lymph nodes (levels 1–3) and invasive breast cancer that has spread beyond the breast and nearby lymph nodes to other organs of the body (level 4) [1]. However, differentiation between these stages is sometimes difficult and needs a lot of medical experience. Further, it might be for some treatments necessary to differentiate between certain stages whereas the distinction between other stages can be neglected for those treatments. Hence, training a classifier system for such tasks should take this expert knowledge into account.

We suggest to modify the class dissimilarities $\delta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_j\right)$ for a respective knowledge integration. One simple choice is to take this dissimilarity as a positive definite bi-linear form

$$\delta_\mathbf{K}\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_j\right) = \left(\mathbf{c}_\mathbf{v} - \mathbf{y}_j\right)^T \mathbf{K}\left(\mathbf{c}_\mathbf{v} - \mathbf{y}_j\right) \tag{8}$$

with the a positive definite expert *knowledge matrix* $\mathbf{K}$. For example, $\mathbf{K}^{-1}$ could be the matrix of anti-/correlations such that the dissimilarity (8) becomes the Mahalanobis distance. Another possibility of knowledge integration is a class weighting related to their importance according to a weighted Euclidean distance

$$\delta_\beta\left(\mathbf{c}_\mathbf{v}, \mathbf{y}_j\right) = \sum_{k=1}^{N_C} \beta_k \left(\left[\mathbf{c}_\mathbf{v}\right]_k - \left[\mathbf{y}_j\right]_k\right)^2 \tag{9}$$

with weights $\beta_k > 0$. In the later application we also consider another type of knowledge integration: As mentioned above, the labels of training data samples may suffer from an uncertainty in the expert labeling. We consider the conditional probability $p(l|k)$ that a data vector assigned to class k could belong to class l. In the following we denote such a case as *failure event* (fe). We collect this information in an uncertainty matrix $\mathbf{U} \in [0,1]^{N_C \times N_C}$ with $U_{k,l} = p(l|k)$ such that $\sum_l U_{k,l} = 1$. Further, we suppose that the diagonal elements are maximum, i.e. $U_{k,k} > U_{k,l}\ \forall l \neq k$. Under this assumption and having non-vanishing off-diagonal elements, a failure event should be less contribute to an error criterion compared to the case that $\mathbf{U}$ is diagonal. For this purpose, we introduce the *unification dissimilarity* for data and prototype label vectors $\mathbf{c_v}$ and $\mathbf{y}_j$

$$D_{k,l} = \left(\frac{[\mathbf{c_v}]_k + [\mathbf{c_v}]_l}{2} - \frac{[\mathbf{y}_j]_k + [\mathbf{y}_j]_l}{2}\right)^2 \tag{10}$$

with respect to the classes k and l. All values $D_{k,l}$ form the unification dissimilarity matrix $\mathbf{D}$. There, the measure $D_{k,l}$ describes the deviation of class vector entries if the classes k and l would be merged, i.e. we take these as a unification.

Both aspects, uncertainty and unification distance are combined in the class dissimilarity measure

$$\delta_{\mathbf{U}}(\mathbf{c_v}, \mathbf{y}_j) = Fr(\mathbf{U} \circ \mathbf{D}) \tag{11}$$

where $\mathbf{U} \circ \mathbf{D}$ is the Hadamard product and $Fr(\cdot)$ denotes the Frobenius norm. Obviously, $\delta_{\mathbf{U}}(\mathbf{c_v}, \mathbf{y}_j)$ is not a metric but still a dissimilarity measure [9].

We emphasize at this point that the uncertainty matrix does not explicitly code the classification goal rather than structural information and expert knowledge about classification failures. Of course, other choices of combinations are possible as well as other structures of expert knowledge.

4 Application

We demonstrate the integration of expert knowledge for a classification problem in the field of neurology. In particular, we consider a data set of $M = 122$ data vectors describing seven neurophysiological parameters in the brain from patients suffering from Morbus Wilson and volunteers [2, 6]. Yet, the volunteers sometimes show neurophysiological impairment symptoms but with lower degree and complexity, which may be symptoms of other diseases. Morbus Wilson is a copper metabolism disease which leads to an accumulation of copper in the brain causing motoric disturbances as well as neurophysiological impairments. According to a clinical scheme suggested by Konovalov, patients can be divided into two main groups: neurological (N) and non-neurological (NN) manifestation [4]. These groups are further differentiated. The neurological group contains the subgroups of pseudo-sclerotic (PS), pseudo-parkinsonian (PP) and merged type (MT). The non-neurological group includes the hepatic type (HT) and the asymptomatic type (AT) [5]. Yet, a clinical (expert) distinction between these subgroups is difficult and requires a strong medical experience. During the course of the disease the non-neurological becomes manifest in the neurological state. Medical treatment may slow this process down and reduces the symptoms. Depending on the impairment level and the respective pharmaceutic dose rate the treatment causes side effects and could also be expensive. Therefore, a precise classification is demanded. Different physical examinations are usually

		PS	PP	MT	HT	AT	V		PS	PP	MT	HT	AT	V
FSNG	PS	16	8	6	2	2	0	FSNG with β	17	2	12	0	2	1
	PA	0	10	1	0	2	1		0	11	1	0	1	1
	MT	2	1	4	0	0	1		0	1	6	0	0	1
	HT	0	6	0	1	0	3		0	6	0	1	0	3
	AT	0	0	0	0	4	4		0	0	0	0	4	4
	V	0	15	1	0	10	22		0	14	1	0	8	25
FSNG with **U**	PS	16	4	11	1	2	0	**U**	0.60	0.15	0.25	0	0	0
	PP	0	1	1	5	2	5		0.10	0.60	0.30	0	0	0
	MT	0	0	6	1	0	1		0.25	0.25	0.50	0	0	0
	HT	1	0	0	4	0	5		0	0	0	0.80	0.15	0.05
	AT	0	0	0	0	4	4		0	0	0	0.15	0.60	0.25
	V	0	0	1	1	10	36		0	0	0	0	0.05	0.95

Table 1: Results of the FSNG using different levels of expert knowledge: confusion matrices for the standard FSNG (top - left), FSNG with class weighting (top right) and expert knowledge intergartion (down - left) and the uncertainty matrix **U** .

applied including genetic analysis, fine-motoric and neurophysiological tests, and other. The results are condensed in the expert diagnosis by the medical doctor.

The task here is, to classify the patients according to the Konovalov-scheme only on the basis of their neurophysiological data $\mathbf{v} \in \mathbb{R}^7$ [4]. Earlier investigations have shown that a neurophysiological persistent manifestation takes place [6]. However, it is not clear whether a precise classification based on these data is possible. Thereby, it is of great importance to differentiate at least between the main types (N, NN) and volunteers. The training data are labeled according to the Konovalov-scheme and then unary coded by class label vectors $\mathbf{c_v}$ with the class sequence (PS,PP,MT,HT,AT,V). Thus the first three entries describe the neurological class followed by the two non-neurological subtypes and the volunteer class.

The structural medical expert knowledge is available (as common sense) for failure events in clinical classification of patients. It is fed into the uncertainty matrix **U** in this way that the non-diagonal elements in each row (diagnosis) of **U** describe the probability for detecting the respective diagnosis (column) instead. The resulting matrix is displayed in Tab. 4. The class dissimilarity is determined according to (11). Otherwise, if less information should be used at least a weighting of the single subtypes is possible: In that case main weight should be given to detect the volunteer group to prevent an unnecessary treatment. According to medical expertise, we used the class weighting vector $\beta = (2, 2, 2, 1, 1, 10)^T$ with class dissimilarity measure (9).

The classification results are generated using the FSNG algorithm to deal also with fuzzy decisions taking $N = 11$ prototypes and identical initialization for all experiments. The resulting confusion matrices are depicted in Tab. 4 in comparison to the results for the standard Euclidean distance as class dissimilarity, which does not include any expert knowledge. We observe for the standard case many misclassifications for volunteers. Further, violations of the mapping to the neurological and the non-neurological occur. Integration of expert knowledge reduces these effects. When weighting of the classes is applied, a small improvement is achieved. However, the remaining violations are remarkable. Integration of stronger expert knowledge about the uncertainty of medical doctors classification leads to a substantial improvement. In particular, the volunteer group is clearly separated from the neurological class. A remaining violation is due to the asymptomatic subtype. This effect is in agreement with clinical findings, because the AT-group usually show very weak symptoms. Additionally, the separation between the neurological and the non-neurological types is also improved. Yet, the confusions within these groups are not solved reflecting the structural expert knowledge. In conclusion we can state that a classification of Wilson's disease types based on neurophysiological measurements is possible if structural expert knowledge is used for the task specific classification scheme. Without this expert information a precise classification is difficult at least.

5 Concluding Remarks

In this paper we propose the integration of structural expert knowledge into classification schemes which make use from class dissimilarity measures. For those systems the expert information can be plugged into this class dissimilarity. We show in an exemplary application of classification of Wilson's disease based on neurophysiological data, how structural knowledge improves the classification results. As classification scheme we used the fuzzy supervised neural gas model, which easily allows such an integration. However, on the basis of this idea we recommend the integration of expert knowledge also for other classification schemes like fuzzy supervised SOM [7] and tasks, if available.

References

[1] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, and A. Trotti, editors. *AJCC cancer staging manual.* Springer, 2010.

[2] P. Günther, T. Villmann, and W. Hermann. Event related potentials and cognitive evaluation in Wilson's disease with and without neurological manifestation. *Journal of Neurological Sciences [Turkish]*, 28(1):79–85, 2011.

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

[4] W. Hermann, P. Günther, A. Wagner, and T. Villmann. Klassifikation des Morbus Wilson auf der Basis neurophysiologischer Parameter. *Der Nervenarzt*, 76:733–739, 2005.

[5] W. Hermann, T. Villmann, F. Grahmann, H. Kühn, and A. Wagner. Investigation of fine motoric disturbances in Wilson's disease. *Neurological Sciences*, 23(6):279–285, 2003.

[6] W. Hermann, T. Villmann, and A. Wagner. Elektrophysiologisches Schädigungsprofil von Patienten mit einem Morbus Wilson'. *Der Nervenarzt*, 74(10):881–887, 2003.

[7] M. Kästner and T. Villmann. Fuzzy supervised neural gas for semi-supervised vector quantization – theoretical aspects. *Machine Learning Reports*, 5(MLR-02-2011):1–16, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_02_2011.pdf.

[8] 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.

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

[10] F.-M. Schleif, T. Villmann, B. Hammer, and P. Schneider. Efficient kernelized prototype based classification. *International Journal of Neural Systems*, 21(6):443–457, 2011.

[11] P. Schneider, B. Hammer, and M. Biehl. Distance learning in discriminative vector quantization. *Neural Computation*, 21:2942–2969, 2009.

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

[13] T. Villmann and M. Kästner. Fuzzy supervised neural gas with sparsity constraint. *Machine Learning Reports*, 5(MLR-05-2011):17–20, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_05_2011.pdf.