ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Reliable Patient Classification in Case of Uncertain Class Labels Using a Cross-Entropy Approach

A. Villmann[1], M. Kaden[2], S. Saralajew[3], W. Hermann[4], and T. Villmann[2]

1- Berufliches Schulzentrum Döbeln-Mittweida, Dep. Mittweida
Mittweida - Germany

2- University of Applied Sciences Mittweida
Saxony Institute for Comp. Intelligence and Machine Learning
Mittweida - Germany

3- Dr. Ing. h.c. F. Porsche AG
Electrical/Electronics Driver Assistance Platform/Systems
Weissach - Germany

4- Spital Langenthal - Neurologie SRO AG
Langenthal - Switzerland

Abstract. Classification learning crucially depends on the correct label information in training data. We consider the problem that a respective uncertainty can neither be neglected nor it can be approximated by a statistical model. In the proposed approach each training data is equipped with a certainty value reflecting the probability of the label correctness. This information is used in the learning process for the classifier. For this purpose, we adopt the cross-entropy cost function from deep learning for a modified learning vector quantization model. We show the usefulness of this knowledge integration in medical diagnostic data analysis for detection of Wilson's disease as an example.

1 Introduction

Classification learning in medicine and bioinformatics becomes more and more successful due to the availability of advanced classifier models like deep architectures together with sophisticated training procedures [5, 13]. Yet, the success of classification learning crucially depends on the training data. One important aspect is the reliability of the data labels, i.e. the correctness of the assignment of the training data to the available classes [4]. Yet, for many application areas this is a serious problem, like in remote sensing image analysis. Here image pixel cover a certain ground area which can contain a mixture of soil or vegetation such that a specific class assignment of a single pixel might be crucial [17]. One solution for this problem could be multi-class models [18]. Often, general classification uncertainties for training data are handled assuming a statistical model for the label noise [2, 11].

In medical data analysis the labels of diagnostic data are the respective medical diagnoses made by the doctor. However, often multiple diagnoses are given or the medical indication only allows a weak decision. For example, psychotherapeutic diagnoses frequently are equipped with a high degree of uncertainty regarding a specific mental disorder. This kind of uncertainty usually does not follow a statistical distribution assumption and, therefore, has to be carefully distinguished from the above mentioned multi-class problem [15]. More likely, the possible doubt of medical doctors is case specific depending on the symptoms and results of the medical examinations.

Otherwise, medical doctors usually have a keen sense, whether a certain diagnosis for a patient is equipped with high vagueness or not. Sometimes, this additional knowledge is available for the trainings data but then usually ignored by the learning system because of the intractability as a statistical model.

In this contribution we propose an approach to deal with situations where this additional knowledge regarding the data specific label uncertainty is available for training data. Additionally, we are faced with the problem that only a few data are available for training as it is frequently the case for medical problems. Yet, this causes difficulties in deep learning, particularly, successful pretraining techniques like auto-encoders for deep networks cannot be applied for those tasks [1, 9]. Thus, we have to concentrate on classifier methods like learning vector quantizers (LVQ,[10]), which can be applied also for those few data. Moreover, we modify the generalized LVQ (GLVQ,[12]) taking the advantage of the cross-entropy cost function. As it was shown in [14], combination of ideas from deep architectures and easy-to-interpret LVQ models generally seems to be a promising way to deal with specific problems. The cross-entropy is usually applied in deep learning architectures, due to its excellent performance behavior for gradient descent learning techniques [5]. Yet, the cross-entropy requires a probabilistic decision model. For this purpose, we tackle the label uncertainty as a statistical model for the binary problem of correct or incorrect class labeling specifically for each data.

The medical problem, for which we apply our method, is the detection of the rare Wilson's disease (WD,[6]) based on medical examination results. These data labels are provided by medical experts together with a value indicating their certainty regarding the diagnosis.

2 Classification Learning with Uncertainties in GLVQ Using Cross-entropy Costs

2.1 Basic GLVQ

GLVQ belongs to the set of prototype based classifiers, i.e. prototypes (reference vectors) serve as representatives for the class distribution. For training vector data $\mathbf{x} \in X \subseteq \mathbb{R}^n$ with class labels $c(\mathbf{x}) \in \mathscr{C} = \{1, \ldots, C\}$ a set of prototypes $W = \{\mathbf{w}_k\}_{k=1}^{M}$ is distributed in the data space X. Thereby, each prototype is assigned to a class $c(\mathbf{w}_k) \in \mathscr{C}$ such that each class is represented by at least one prototype. In the application mode an unknown data vector $\mathbf{x}$ is assigned to the data class $c(\mathbf{w}_s)$ according to the winner-takes-all (WTA) rule

$$s(\mathbf{x}) = \mathrm{argmin}_k d(\mathbf{x}, \mathbf{w}_k) \tag{1}$$

determining the best-matching prototype $\mathbf{w}_s$ with minimal dissimilarity measure $d(\mathbf{x}, \mathbf{w}_k)$, frequently chosen as the Euclidean distance. The set $R_{\mathbf{w}_k} = \{\mathbf{x} \in X | k = s(\mathbf{x})\}$ is denoted as the *receptive field* of prototype $\mathbf{w}_k$.

Following [12], learning in GLVQ takes place as a stochastic gradient descent for the cost function

$$E_{GLVQ}\left(X,W\right)=\sum_{\mathbf{x}\in X}\varphi_{\theta}\left(\mu\left(\mathbf{x},W\right)\right)\tag{2}$$

with local costs $\varphi_{\theta}\left(\mu\left(\mathbf{x},W,d\right)\right)$ approximating the local classification error. Here, $\varphi_{\theta}\left(z\right)=1/\left(1+\exp\left(z/\theta\right)\right)$ is the sigmoid transfer function such that in the limit $\theta\searrow 0$ the transfer function becomes the Heaviside function. The argument

$$\mu\left(\mathbf{x},W\right)=\frac{d^{+}\left(\mathbf{x}\right)-d^{-}\left(\mathbf{x}\right)}{d^{+}\left(\mathbf{x}\right)+d^{-}\left(\mathbf{x}\right)}\tag{3}$$

is denotes as the classifier function, where $d^{+}\left(\mathbf{x}\right)=d\left(\mathbf{x},\mathbf{w}^{+}\right)$ is the dissimilarity of the best matching prototype $\mathbf{w}^{+}=\mathbf{w}_{s^{+}}$ with correct class assignment, i.e

$$s^{+}=\operatorname{argmin}_{k=1\ldots M|c(\mathbf{w}_{k})=c(\mathbf{x})}d\left(\mathbf{x},\mathbf{w}_{k}\right)\tag{4}$$

and $\mathbf{w}^{-}=\mathbf{w}_{s^{-}}$ is, in analogy, the best matching prototype for all incorrect classes. Stochastic gradient descent learning of $E_{GLVQ}\left(X,W,d,\theta\right)$ takes place as prototype updates according to

$$\Delta\mathbf{w}^{\pm}\propto\frac{\partial\varphi_{\theta}\left(\mu\right)}{\partial\mu}\cdot\nabla_{\mathbf{w}^{\pm}}\left(\mu\right)\text{ with }\nabla_{\mathbf{w}^{\pm}}\left(\mu\right)=\frac{\partial\mu}{\partial d^{\pm}}\frac{\partial d^{\pm}}{\partial\mathbf{w}^{\pm}}\tag{5}$$

calculated pursuant to the chain rule for derivatives.

2.2 Incorporation of Label Uncertainty of Training Data into GLVQ

As motivated in the introduction we now turnover to the incorporation of label uncertainty into the GLVQ model assuming a certainty value $\zeta\left(\mathbf{x}\right)\in\left[0,1\right]$ for correct labeling of a training data $\mathbf{x}$. The higher the value $\zeta\left(\mathbf{x}\right)$ the higher is the certainty that the data label is true. This value has to be delivered by the users who generated the training data evaluating their uncertainty regarding their classification decision. Appropriately, we can interpret the value $\zeta_{W}\left(\mathbf{x}\right)=\varphi_{\theta}\left(-\mu\left(\mathbf{x},W\right)\right)\in\left[0,1\right]$ as the certainty of the GLVQ regarding a correct classification. This choice is motivated by the idea of reject options for vague decisions based on the classifier value $\mu\left(\mathbf{x},W\right)$ [3, 16]. Hence, both certainty values $\zeta\left(\mathbf{x}\right)$ and $\zeta_{W}\left(\mathbf{x}\right)$ can be seen as probabilities for correct labeling and classification, respectively, whereas $1-\zeta\left(\mathbf{x}\right)$ and $1-\zeta_{W}\left(\mathbf{x}\right)$ are the converse probabilities for incorrect labeling and classification. Hence, we consider a probabilistic binary problem now.

Thus, we can reformulate the training task for classification learning in this way that we want to achieve a high compliance between the certainty values for correct labeling/classification of the data. Due to the probabilistic view, we can adopt the idea of deep learning cost function for our binary problem resulting in the (negative) local cross-entropy

$$Cr\left(\mathbf{x},W\right)=-\zeta\left(\mathbf{x}\right)\cdot\ln\left(\zeta_{W}\left(\mathbf{x}\right)\right)-\left(1-\zeta\left(\mathbf{x}\right)\right)\cdot\ln\left(1-\zeta_{W}\left(\mathbf{x}\right)\right)$$

for each data $\mathbf{x}$. Thus we obtain the new cost function

$$E_{CrGLVQ-U}\left(X,W\right)=\sum_{\mathbf{x}\in X}Cr\left(\mathbf{x},W,\varphi_{\theta}\right)\tag{6}$$

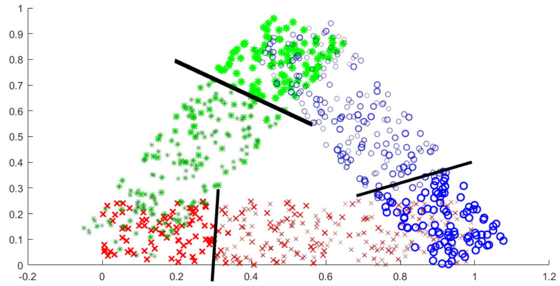

Fig. 1: Triangle dataset - three overlapping classes. Bold symbols indicate a label certainty $\zeta(\mathbf{x}) = 1$ whereas for the other data $\zeta(\mathbf{x}) < 1$ is valid. The different certainty areas within the classes are separated by black lines.

to be minimized by stochastic gradient descent learning. The respective prototype learning rule is obtained as

$$\Delta\mathbf{w}^{\pm} \propto \frac{\partial Cr(\mathbf{x}, W)}{\partial \zeta_W(\mathbf{x})} \cdot \frac{\partial \zeta_W(\mathbf{x})}{\partial \mu} \cdot \nabla_{\mathbf{w}^{\pm}}(\mu) \tag{7}$$

which can be calculated using the relations

$$\frac{\partial Cr(\mathbf{x}, W)}{\partial \zeta_W(\mathbf{x})} = -\frac{\zeta(\mathbf{x})}{\zeta_W(\mathbf{x})} + \frac{1-\zeta(\mathbf{x})}{1-\zeta_W(\mathbf{x})} \text{ and } \frac{\partial \zeta_W(\mathbf{x})}{\partial \mu} = -\frac{\partial \varphi_\theta(\mu)}{\partial \mu}$$

for the derivatives. The resulting GLVQ variant is denoted as *Cross-entropy GLVQ with Uncertainty* (CrGLVQ-U).

3 Experiments

We report the results of two experiments we conducted. The first experiment is a two-dimensional toy example to illustrate the method. The second one is the classification of patients regarding their electro-physiological impairment profiles in case of Wilson's disease detection.

Toy Example - the Triangle Dataset The *triangle dataset* consists of 2D-data belonging to three overlapping classes, see Fig.1. We trained a GLVQ as well as a CrGLVQ-U with one prototype for each class. The certainty value in regions with increased uncertainty was set randomly as $\zeta(\mathbf{x}) \in [0.5, 1]$. However, only CRGLVQ-U uses this additional information. The resulting prototype configurations are visualized in Fig.2. We observe that the prototypes of CrGLVQ-U try to be responsible for areas with high label certainty whereas the GLVQ-prototypes are located more in the class centers despite the weak certainty information in this region. To reflect this behavior precisely, we introduce the quantity

$$\zeta_{\mathbf{w}}(X) = \frac{\#\left\{\mathbf{x} \in X | \mathbf{w} = \mathbf{w}_{s(\mathbf{x})} \wedge c(\mathbf{x}) = c\left(\mathbf{w}_{s(\mathbf{x})}\right)\right\}}{\#\left\{\mathbf{x} \in X | \mathbf{w} = \mathbf{w}_{s(\mathbf{x})}\right\}}$$

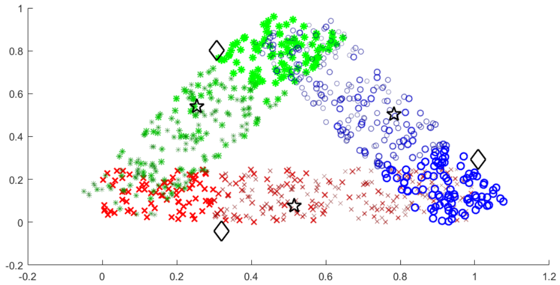

Fig. 2: Learned prototypes for standard GLVQ (★) and CrGLVQ-U (◊). The prototypes for the CrGLVQ-U moved to the class areas with high certainty value to ensure a secure classification whereas standard GLVQ-prototypes do not benefit from this information and, therefore, are located in regions with high uncertainty.

denoted as the *classification certainty* of the prototype $\mathbf{w}$ regarding the prototype class $c(\mathbf{w})$. Hence, the classification certainty of $\mathbf{w}$ is the fraction of data points in the receptive field $R_{\mathbf{w}}$ belonging to the same class like $\mathbf{w}$. The overall model classification certainty is the average $\zeta(X,W) = \frac{1}{|W|}\sum_{\mathbf{w}\in W}(\zeta_{\mathbf{w}}(X))$. For the GLVQ we obtain $\zeta(X,W) = 0.7357$ whereas the CrGLVQ-U yields $\zeta(X,W) = 1$ indicating a clear improvement of he classification certainty.

The Wilson's Disease Dataset We apply the model to a real medical dataset for patients suffering from Wilson disease (WD) (with uncertainty) or being considered as healthy. WD is a copper metabolism disturbance leading to neurological impairments. The initial non-neurological phase is followed by a neurological phase with increasing neurological deficits. A precise WD-diagnosis is difficult and costly. One cheaper alternative is to consider electro-physiologic impairment profiles (EIP) consisting of vector of electro-physiological stimulus responses [8]. Yet, just considering those profiles gives increased uncertainty regarding a final diagnoses - particularly the non-neurological phase has high clinical uncertainty [7]. Our dataset consists of six-dimensional EIP-data 77 WD-patients, 59 diagnosed as neurological patients, i.e. having a certainty value for WD-diagnosis of 1 whereas 18 patients are diagnosed as non-neurological patients giving a certainty value 0.7 for WD-diagnosis. Additionally, the EIP from 48 volunteers are available (certainty 0.999 - reflecting WD-prevalence). We applied GLVQ and CrGLVQ-U with one prototype per class, the results are given for a five-fold cross-validation. For the standard GLVQ we obtained an averaged model certainty of $\zeta(X,W) = 0.84$ (± 0.054) whereas CrGLVQ-U achieved a certainty $\zeta(X,W) = 0.89$ (± 0.015). Hence, the incorporation of label certainty information in learning leads to a better classification certainty.

4 Conclusion

In this contribution we consider the incorporation of label certainty for training data into the learning strategy. For this purpose, we modified the standard

GLVQ to utilize this additional knowledge during training. Particularly, we introduced certainty values for labels to obtain a probabilistic model, which enables to apply the cross-entropy cost function known from deep learning. Thus, a better model classification certainty is obtained for the trained classifier.

References

[1] Y. Bengio. *Neural Networks: Tricks of the Trade*, chapter Practical Recommandations for gradient-based training of deep architectures, p. 437–478. Springer, Berlin, 2012.

[2] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2):2780–2824, 2016.

[3] L. Fischer, B. Hammer, and H. Wersing. Efficient rejection strategies for prototype-based classification. *Neurocomputing*, 169:334–342, 2015.

[4] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[6] W. Hermann, H. Barthel, S. Hesse, F. Grahmann, H.-J. Kühn, A. Wagner, and T. Villmann. Comparison of clinical types of Wilson's disease and glucose metabolism in extrapyramidal motor brain regions. *Journal of Neurology*, 249(7):896–901, 2002.

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

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

[9] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(7):5, 2006.

[10] T. Kohonen. Learning Vector Quantization. *Neural Networks*, 1(Suppl. 1):303, 1988.

[11] T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

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

[13] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[14] T. Villmann, M. Biehl, A. Villmann, and S. Saralajew. Fusion of deep learning architectures, multilayer feedforward networks and learning vector quantizers for deep classification learning. In *Proceedings of the 12th Workshop on Self-Organizing Maps and Learning Vector Quantization (WSOM2017+)*, pages 248–255. IEEE Press, 2017.

[15] T. Villmann, G. Blaser, A. Krner, and C. Albani. Relevanzlernen und statistische Diskriminanzverfahren zur ICD-10 Klassifizierung von SCL90-Patienten-Profilen bei Therapiebeginn. In G. Plttner, editor, *Aktuelle Entwicklungen in der Psychotherapieforschung*, pages 99–118. Leipziger Universittsverlag, Leipzig, Germany, 2004.

[16] T. Villmann, M. Kaden, A. Bohnsack, S. Saralajew, J.-M. Villmann, T. Drogies, and B. Hammer. Self-adjusting reject options in prototype based classification. In E. Merényi, M. Mendenhall, and P. O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proc. of 11th Int. Workshop WSOM 2016*, vol. 428 of *Advances in Intelligent Systems and Computing*, p. 269–279, Berlin-Heidelberg, 2016. Springer.

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

[18] M. Zhang and Z. Zhou. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.