The LVQ-based Counter Propagation Network – an Interpretable Information Bottleneck Approach

M. Kaden, R. Schubert, M. Mohannazadeh Bakhtiari, L. Schwarz, and T. Villmann *

University of Applied Sciences Mittweida, Saxon Institute for Comp. Intelligence and Machine Learning, Mittweida - Germany

Abstract. In this paper we present a realization of the informationbottleneck-paradigm by means of an improved counter propagation network. It combines an unsupervised vector quantizer for data compression with a subsequent supervised learning vector quantization model. The approach is mathematically justified and yields an interpretable model for classification under the constraint of data compression, which is not longer independently learned from the classification task.

1 Introduction and Motivation

The processing of complex data is of increasing importance because the availability of more and more complex data pervades many branches in industry and research. For example, high-resolution spectrometers in medicine and bioinformatics deliver huge-dimensional data records. In astrophysics terabytes of data arrive daily and often require instantaneous decisions regarding storage or/and further processing. Frequently, the data can be kept as compressed data usually accompanied by an information loss. In fact data compression may destroy valuable information needed later for data evaluation. In information theory this problem is known as the information bottleneck principle (IBP) [24].

In this paper we address this problem in context of interpretable classification learning, i.e. we propose a vector quantization based approach to realize an information bottleneck network for classification. The starting point is the heuristically motivated counter propagation network (CPN) introduced by R. HECHT-NIELSEN [10, 11]. The CPN network consists of two steps: First, the data are sluiced through a self-organizing map (SOM) network for information compression and clustering [14]. Second, a perceptron layer evaluates the SOMresponse for classification or regression learning. Recently, the idea of this vector quantization combination with perceptron layers was renewed by C. RUDIN for deep multi-layer perceptrons to achieve better robustness [6]. Yet, the CPN approach offers at least two drawbacks: Although the SOM vector quantizer frequently yields a surprisingly good data representation, a mathematical guarantee for optimum data coding can not be given [7]. Further, the unsupervised SOM training is independent from the subsequent supervised perceptron training. This motivates our approach: We replace the SOM-layer by a neural gas vector quantizer, which is mathematically justified and constitutes an information optimum data compression for high-dimensional data [17]. Further, we substitute the perceptron layer by a learning vector quantization (LVQ) layer evaluating the vector quantizer response for interpretable classification learning.

 $^{^{*}\}mathrm{M.K.,\,M.M.B.}$ and D.S. were supported by grants of the European Social Fund (ESF) for a Young Researcher Group 'MaLeKITA' and a PhD grant.

Both layers are coupled via gradient training, i.e. the neural gas layer receives information from the subsequent classification layer during training. In this way, the compression process is influenced by the subsequent classification task realizing the IBP.

The presented approach is also in relation to [4], where a combined compression and classification method by learning vector quantization is considered. Yet, the classification decision there is based on majority vote regarding the receptive fields by a *c*-means vector quantizer [16]. In [15] as well as in [18, 19], classification learning based on quantizer codebooks is investigated by information loss minimization using Gaussian mixture models to classify. For optimization they use an alternating update heuristics.

The paper is structured as follows: First we briefly present the original counter propagation network. Thereafter we explain the improvements provided by our approach. Numerical simulations and concluding remarks are the final part of this contribution.

2 The Original Counter Propagation Network

As already mentioned, the idea of a counter propagation network (CPN) is to combine a SOM vector quantizer for the representation of data $X \subseteq \mathbb{R}^n$ with a perceptron layer for supervised learning of class labels $c(\mathbf{x}) \in \mathscr{C} = \{1, \ldots, C\}$. The perceptron layer is also denoted as Grossberg-layer, as defined in [9], in this context and was later also used in adaptive-resonance-theory (ART) networks [5]. We will denote the first layer as the *sensoric vector quantization layer* and the second layer as the *classification layer* to emphasize the more general context later in this paper.

The SOM supposes that the *sensoric* prototype set $W \subseteq \mathbb{R}^n$ is related to an external *sensoric grid* $\mathscr{S} \subset \mathbb{R}^p$ [14], inspired by the cortical areas in the human brain processing all sensoric information. Thus, the projection dimension p is chosen as p = 2. Here, we assume K prototypes $\mathbf{w}_k \in W$ where the index $\mathbf{r}(k) \in \mathscr{S}$ refers to a location in the sensoric grid \mathscr{S} and $k(\mathbf{r}) \in \{1, \ldots, K\}$ returns the respective index in W. For a given input $\mathbf{x} \in X$ the most appropriate prototype is determined by the winner-takes-all (WTA) rule

$$s\left(\mathbf{x}\right) = \operatorname{argmin}_{k(\mathbf{r})|\mathbf{r}\in\mathscr{S}} d_E\left(\mathbf{x}, \mathbf{w}_k\right) \tag{1}$$

with $d_E(\mathbf{x}, \mathbf{w}_k)$ being the (squared) Euclidean distance realizing the Hebbian excitation [20]. We define a response vector $\boldsymbol{\xi}(\mathbf{x}, W)$ with $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1, \dots, \xi_K)^T \in \Xi \subseteq \mathbb{R}^K$ with Ξ denoting the sensoric response field. The WTA-rule delivers

$$\xi_{k(\mathbf{r})}(\mathbf{x}, W) = \begin{cases} 1 & k = s(\mathbf{x}) \\ 0 & k \neq s(\mathbf{x}) \end{cases}$$
(2)

as stimulus response. We can interpret this mapping $\mathbf{x} \mapsto \boldsymbol{\xi}(\mathbf{x}, W)$ as an information compressing mapping realized by a vector quantizer. The prototype adaptation is

$$\Delta \mathbf{w}_{k} \propto -h_{\lambda}^{SOM}\left(\mathbf{r}\left(k\right), \mathbf{x}, W\right) \cdot \frac{\partial d\left(\mathbf{x}, \mathbf{w}_{k}\right)}{\partial \mathbf{w}_{k}}$$
(3)

where

$$h_{\lambda}^{SOM}\left(\mathbf{r}\left(k\right),\mathbf{x},W\right) = \exp\left(\frac{-\left\|\mathbf{r}\left(s\left(\mathbf{x}\right)\right) - \mathbf{r}\left(k\right)\right\|_{\mathscr{S}}}{2\sigma^{2}}\right)$$

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

is the grid dependent neighborhood function with neighborhood range σ . The perceptron layer consists of a single linear perceptron realized as

$$y\left(\mathbf{x}\right) = \boldsymbol{\omega}^{T} \cdot \boldsymbol{\xi}\left(\mathbf{x}, W\right) \tag{4}$$

with an adjustable perceptron weight vector $\boldsymbol{\omega}$ to achieve a good class label prediction $c(\mathbf{x}) = y(\boldsymbol{\xi}(\mathbf{x}, W))$. We can summarize the CPN scheme as

$$X \xrightarrow{P(X)}_{\text{SOM}} W \xrightarrow{\boldsymbol{\xi}(\mathbf{x},W)}_{\text{crisp}} \Xi \xrightarrow{y(\boldsymbol{\xi}(\mathbf{x},W))}_{\text{perceptron layer}} \mathscr{C}$$
(5)

where P(X) is the data density. Training in CPN takes place in two phases: First, the VQ-layer (SOM) is trained in an unsupervised manner. Second, the perceptron weight vector $\boldsymbol{\omega}$ is adapted by supervised learning. Thus, the SOM layer yields a grouping of data whereas the perceptron learns to interpret this grouping for classification learning.

At least two improvements are discussed in the community so far:

- 1. Instead of only one perceptron, class-wise non-linear perceptrons are taken [8], i.e. $y_k(\mathbf{x}) = f(\boldsymbol{\omega}_k^T \cdot \boldsymbol{\xi}(\mathbf{x}, W) \beta_k)$ with biases β_k and an activation function f frequently taken as sigmoid.
- 2. Several authors suggested to relax the WTA-rule taking more than a single winning unit [8, 12]. For this purpose, N units of the SOM layer surrounding the winning unit $\mathbf{r}(s(\mathbf{x}))$ in \mathscr{S} are taken with $\xi_{k(\mathbf{r})}(\mathbf{x}, W) = \frac{1}{N}$ for all these including the winning neuron $\mathbf{r}(s(\mathbf{x}))$.

The CPN approach frequently works very successful although being simple [3, 8, 26]. One can see this also as an early approach to incorporate prototype layers in multi-layer perceptrons as recently discussed in [6]. Yet, the SOM-training takes place independently from the subsequent classification task and, hence, might be suboptimal for the later classification learning. Further, the original SOM does not optimize any cost function such that mathematical guarantees for data grouping behavior are given [7]. Therefore, we propose new variants and extension of this basic CPN.

3 Proposed Modifications of the CPN

An alternative to SOM is the neural gas vector quantizer (NG) [17]. For NG, neighborhood of the prototypes is given implicitly by the exponential winning-rank-function of the prototype \mathbf{w}_k

$$h_{\lambda}^{NG}(k, \mathbf{x}, W) = \exp\left(\frac{-\sum_{j} H\left(d\left(\mathbf{x}, \mathbf{w}_{k}\right) - d\left(\mathbf{x}, \mathbf{w}_{j}\right)\right)}{\lambda}\right) \text{ with } H\left(z\right) = \begin{cases} 1 & \text{for } z > 0\\ 0 & \text{elsewere} \end{cases}$$
(6)

being the Heaviside function. The prototype dynamic is as (3) for SOM replacing there the SOM-neighborhood function by $h_{\lambda}^{NG}(k, \mathbf{x}, W)$ and reads as

$$\Delta \mathbf{w}_{k} \propto -\frac{\partial \boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right)}{\partial \mathbf{w}_{k}} \stackrel{\text{in prob.}}{=} h_{\lambda}^{NG}\left(k,\mathbf{x},W\right) \cdot \frac{\partial d\left(\mathbf{x},\mathbf{w}_{k}\right)}{\partial \mathbf{w}_{k}}$$
(7)

involving the neighborhood cooperativeness. In fact, it performs a stochastic gradient descent on a cost function, which can be interpreted as the potential function of this gas [17]. Further, according to the magnification property of

NG, resulting from the neighborhood-cooperative learning of NG (7), we can conclude that NG forms an information optimum data representation by the trained prototypes.

Further, we introduce gradual responses as known from Heskes-SOM

$$\xi_k^{NG}(\mathbf{x}, W) = h_\lambda^{NG}(k, \mathbf{x}, W) \cdot d_{\Psi}(\mathbf{x}, \mathbf{w}_k)$$
(8)

for the NG-layer reflecting both the winning rank as well as the dissimilarity $d_{\Psi}(\mathbf{x}, \mathbf{w}_k) = (\Psi(\mathbf{x} - \mathbf{w}_k))$ between the data and sensoric prototypes, where $\Psi \in \mathbb{R}^{n \times n}$ is a mapping matrix. Further, to keep the interpretability, we replace the perceptron classifier by a prototype based learning vector quantization classifier (LVQ, [13]). Taking the sensoric responses $\boldsymbol{\xi}^{NG}$ as input, the LVQ assumes prototypes $\mathcal{W} = \{|\boldsymbol{\omega}_j \in \mathbb{R}^K\}_{j=1}^M$ with class labels $c(\boldsymbol{\omega}_j)$. According to [22, 23], a classifier function is defined as

$$\mu_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\mathcal{W}\right) = \frac{\delta_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\boldsymbol{\omega}^{+}\right) - \delta_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\boldsymbol{\omega}^{-}\right)}{\delta_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\boldsymbol{\omega}^{+}\right) + \delta_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\boldsymbol{\omega}^{-}\right)} \tag{9}$$

with the dissimilarity measure $\delta_{\Omega} \left(\boldsymbol{\xi}^{NG} (\mathbf{x}, W), \boldsymbol{\omega} \right) = \left(\Omega \left(\boldsymbol{\xi}^{NG} (\mathbf{x}, W) - \boldsymbol{\omega} \right) \right)^2$ and the LVQ mapping matrix $\Omega \in \mathbb{R}^{K \times K}$. Thereby, $\boldsymbol{\omega}^+$ is the best matching prototype $\boldsymbol{\omega}_j$ according to the WTA-rule (1) known from SOM with the constraint of class label agreement $c(\boldsymbol{\omega}_j) = c(\mathbf{x})$. Analogously, $\boldsymbol{\omega}^-$ is the best matching prototype among all prototypes responsible for any other classes than $c(\boldsymbol{\xi}(\mathbf{x}, W))$. Thus, μ_{Ω} delivers negative values for correct classification. An unknown data point \mathbf{x} is mapped to to the class $c\left(\boldsymbol{\omega}_{s(\boldsymbol{\xi}^{NG}(\mathbf{x}, W))}\right)$ according to the WTA rule (1). This generalized matrix LVQ (GMLVQ) is mathematically proven to be a large margin classifier and robust while keeping the interpretability according to the prototype reference principle [21, 25]. Its cost function $E_{GMLVQ}(W) = \sum_{\mathbf{x}} E_L(\mathbf{x}, W, W)$ approximates the overall classification error with local errors $E_L(\mathbf{x}, W, W) = f(\mu(\boldsymbol{\xi}(\mathbf{x}, W), W))$. Here f is a differentiable sigmoid transfer function. Learning for the prototypes $\boldsymbol{\omega}_j$ takes place as the stochastic gradient descent by means of the gradients

$$\nabla_{\boldsymbol{\omega}^{+},\boldsymbol{\omega}^{-}} E_{L}\left(\mathbf{x},\mathcal{W},W\right) = \nabla_{\delta_{\boldsymbol{\Omega}}\left(\boldsymbol{\xi},\boldsymbol{\omega}^{\pm}\right)} E_{L}\left(\mathbf{x},\mathcal{W},W\right) \cdot \frac{\partial \delta_{\boldsymbol{\Omega}}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\boldsymbol{\omega}^{\pm}\right)}{\partial \boldsymbol{\omega}^{\pm}}$$

using the abbreviation

$$\nabla_{\delta_{\Omega}(\boldsymbol{\xi},\boldsymbol{\omega}^{\pm})} E_{L}\left(\mathbf{x},\mathcal{W},W\right) = \frac{\partial f\left(\mu_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\mathcal{W}\right)\right)}{\partial \mu_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\mathcal{W}\right)} \cdot \frac{\partial \mu_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x},W\right),\mathcal{W}\right)}{\partial \delta_{\Omega}\left(\boldsymbol{\xi},\boldsymbol{\omega}^{\pm}\right)}$$

whereas Ω -learning requires the evaluation of $\nabla_{\Omega} E_L(\mathbf{x}, \mathcal{W}, W)$. Yet, considering the gradients

$$\nabla_{\mathbf{w}_{k}} E_{L}\left(\mathbf{x}, \mathcal{W}, W\right) = \nabla_{\delta_{\Omega}\left(\boldsymbol{\xi}, \boldsymbol{\omega}^{\pm}\right)} E_{L}\left(\mathbf{x}, \mathcal{W}, W\right) \cdot \frac{\partial \delta_{\Omega}\left(\boldsymbol{\xi}^{NG}\left(\mathbf{x}, W\right), \boldsymbol{\omega}^{\pm}\right)}{\partial \boldsymbol{\xi}^{NG}\left(\mathbf{x}, W\right)} \cdot \frac{\partial \boldsymbol{\xi}^{NG}\left(\mathbf{x}, W\right)}{\partial \mathbf{w}_{k}}$$

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

we are able to adapt the sensoric prototypes in the VQ-layer in dependence on the classification problem simply making use of the chain rule for derivatives. However, their adaptation dynamic still involves the neighborhood cooperativeness of NG (7) and, therefore, tries to keep information optimum for the VQ-layer realizing the IBP. Further, the gradient $\nabla_{\Psi} E_L(\mathbf{x}, \mathcal{W}, W)$ can be used to adapt the NG-mapping matrix Ψ in dependence of the subsequent classification task.

The modified CPN approach can be summarized in relation to (5) as

$$X \xrightarrow{P(X)}_{\text{NG}} W \xrightarrow{\boldsymbol{\xi}^{NG}(\mathbf{x},W)}_{\text{NG-like}} \Xi \xrightarrow{c\left(\boldsymbol{\omega}_{s\left(\boldsymbol{\xi}^{NG}(\mathbf{x},W)\right)}\right)}_{\text{GMLVQ}} \mathscr{C}$$
(10)

constituting an interpretable realization of the IBP for classification learning.

4 Example Application – Classification of Whisky Spectra

We demonstrate the ability of the proposed model for the classification of whisky Raman spectra, dimensionality n = 401, with respect to the age of the whisky (D1 - three classes) and the kind of wood for the cask (D2 - four classes). The data were used and described in $[1, 2]^1$. Here, we show the performance of the CPN in dependence on separate or combined NG- and GMLVQ learning in CPN as well as in dependence on the compression ratio $c_r = K/n$. The results obtained by three-fold cross validation are depicted in Tab. 1. As we can

	NG+GMLVQ						modified CPN						GMLVQ	
K	2		4		8		2		4		8			
M	1	2	1	2	1	2	1	2	1	2	1	2	1	2
D1	84.7	75.4	85.2	85.8	85.0	82.7	71.7	88.2	98.8	98.8	98.8	98.8	98.8	99.1
D2	70.0	76.8	76.5	82.3	75.1	77.7	78.8	83.8	99.4	99.5	99.3	99.5	99.1	99.4

Table 1: Accuracy results: NG+GMLVQ – NG trained independently from GMLVQ; modified CPN – our approach; GMLVQ – without data compression.

see, combined CPN outperforms the separated learning NG+GMLVQ. Further, moderate compression still leads to satisfying accuracy.

5 Conclusion

We reconsidered CPN as introduced by R. HECHT-NIELSEN. In particular, we replace the SOM-layer by a neural gas layer and the perceptron layer is substituted by a GMLVQ layer. Importantly, learning is realized combining both layers for gradient descent learning applying the chain rule for the derivatives. Thus we obtain a mathematically justified interpretable realization of the information bottleneck principle.

References

- P. Ashok, B. Praveen, and K. Dholakia. Near infrared spectroscopic analysis of single malt Scotch whisky on an optofluidic chip. *Optics Express*, 19(23):1–11, 2011.
 A. Backhaus, P. Ashok, B. Praveen, K. Dholakia, and U. Seiffert. Classifying scotch
- [2] A. Backhaus, P. Ashok, B. Praveen, K. Dholakia, and U. Seiffert. Classifying scotch whisky from near-infrared Raman spectra with a radial basis function network with relevance learning. In M. Verleysen, editor, Pro. of the European Symposium on Artificial Neural Networks and Machine Learning (ESANN), p. 411–416, Brussels, 2012. i6doc.com.

¹Thanks to Prof. U. Seiffert, Fraunhofer IFF Magdeburg, Germany, for sharing the data.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7. Available from http://www.i6doc.com/en/.

- [3] B. Bajželj and V. Drgan. Hepatotoxicity modeling using counter-propagation artificial neural networks: Handling an imbalanced classification problem. Molecules, 25(3):481, 2020.
- [4] J. S. Baras and S. Dey. Combined compression and classification with learning vector quantization. IEEE Transactions on Information Theory, 45:1911-20, 1999.
- [5] G. A. Carpenter and S. Grossberg. The ART of adaptive pattern recognition by a selforganizing neural network. IEEE Computer, 21(3):77-88, 1988.
- [6] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32, pages 8930–8941. Curran Associates, Inc., 2019.
- [7] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: Ordering, convergence properties and energy functions. Biol. Cyb., 67(1):47-55, 1992.
- [8] D. Graupe. Principles of Artificial Neural Networks, vol. 8 of Advanced Series in Circuits and Systems, ch. Counter Propagation, p. 185–201. World Scientific, 3rd ed., 2019.
- [9] S. Grossberg. Some networks that can learn, remember, and reproduce any number of complicated space-time patterns. J. of Mathematics and Mechanics, 19(1):53–91, 1969.
- [10] R. Hecht-Nielsen. Counter progagation networks. Appl. Opt., 26(23):4979-4984, December 1987.
- [11] R. Hecht-Nielsen. Applications of counterpropagation networks. Neural Networks, 1(2):131-139, 1988.
- [12] J. A. Hertz, A. Krogh, and R. G. Palmer. Introduction to the Theory of Neural Computation, volume 1 of Santa Fe Institute Studies in the Sciences of Complexity: Lecture Notes. Addison-Wesley, Redwood City, CA, 1991. T. Kohonen. Learning Vector Quantization. Neural Networks, 1(Suppl. 1):303, 1988.
- [14] T. Kohonen. Self-Organizing Maps, volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).
- [15] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(7):1294-1309, 2009.
- [16] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. IEEE Transactions on Communications, 28:84-95, 1980.
- [17] 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.
- [18] K. L. Oehler and R. M. Gray. Combining image compression and classification using vector quantization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17:461-473, 1995.
- [19] K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, R. A. Olshen, and K. L. Oehler. Bayes risk weighted vector quantization with posterior estimation for image compression and classification. IEEE Trans. on Image Processing, 5(2):347-360, February 1996.
- [20] H. Ritter, T. Martinetz, and K. Schulten. Neural Computation and Self-Organizing Maps: An Introduction. Addison-Wesley, Reading, MA, 1992.
 [21] S. Saralajew, L. Holdijk, and T. Villmann. Fast adversarial robustness certification of
- nearest prototype classifiers for arbitrary seminorms. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), p. 13635-13650. Curran Associates, Inc., 2020.
- [22] 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. Proc. of the 1995 Conference, pages 423-9. Curran Associates, Inc., Cambridge, MA, USA, 1996.
- [23] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. Neural Computation, 21:3532-3561, 2009.
- [24] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing, pages 368-377, 1999.
- [25] J. Zeng, B. Ustun, and C. Rudin. Interpretable classification models for recidivism prediction. J. of the Royal Statistical Society, Series A, 180(0964-1998/17/180000):1-34, 2017.
- [26] J. Zupan, M. Novic, and I. Ruisanchez. Kohonen and counterpropagation artificial neural networks in analytical chemistry. Chemometrics and Intelligent Laboratory Systems, 38:1-23, 1997.