ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Information Theory Related Learning

T. Villmann[1]*, A. Cichocki[2], and J. Principe[3]

[1]University of Applied Sciences Mittweida
Dep. of Mathematics/Natural & Computer Sciences, Computational Intelligence Group
Technikumplatz 17, 09648 Mittweida - GERMANY
[2]Riken Brain Sience Institute
Lab. for Advanced Brain Signal Processing
2-1 Hirosawa,Wako City, Saitama, 351-0198 JAPAN
[3]University of Florida
Computational NeuroEngineering Laboratory
Gainesville, FL 32611 - USA

Abstract. This is the introduction paper to a special session held on ESANN conference 2011. It reviews and highlights recent developments and new direction in information related learning, which is a fastly developing research area. These algorithms are based on the fundamental principles of information theory and relate them implicitly or explicitly to learning algoithms and strategies.

1 Introduction

The amount of available data to be analyzed and processed is continuously increasing. However, normally one is not interested in the data but in the information they contain. Therefore, the need for efficient and reliable algorithms and methods is very actual and increasingly important. Different strategies are possible ranging from unsupervised compression, random selection, projection to supervised feature selection, or shape detection, to name just a few. All these approaches have in common that information contained in the data should be extracted emphasizing different aspects depending on the task, i.e. they realize an information processing system. Hence, these methods are based implicitly or explicitly on information theory and their consequences.

In this paper, we will draw attention to recent developments of this field. Obviously, this is not a complete overview but still picks some interesting new aspects of the ongoing research and remembers also well known facts in that area.

2 Information theoretic learning via statistics and probability

Probability theory and statistics are essentially related to information theory. Statistical distributions like Gaussians and others are directly involved in fundamental theorems of information theory: The second Gibbs theorem about

*corresponding author, *email: thomas.villmann@hs-mittweida.de*

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

maximum Shannon entropy property of normal distributions is the most prominent fact [18]. Similarity measures were related to probability theory [6], leading to divergence measures as generalized distances like the Kullback-Leibler divergence [35]. Trough time, a large variety of divergences were developed and several classes of divergences identified [14],[64]. This research has new focus [39] and involves modern differential-geometrical methods in statistics and probability theory [3] and in statistical learning [21].

Models in machine learning directly include these results into learning algorithms and strategies. One of the earliest approaches connecting statistics, information theory and biologically motivated learning is the perceptron neuron model of neurons [50]. Multilayer networks can be optimized avoiding overtraining using mutual information [17]. Boltzmann networks are derived from information principles of statistical mechanics [1],[42].

Source separation of data channels is based on the statistical deconvolution. Different aspects can be investigated like independent component analysis (ICA) and blind source separation (BSS) maximizing conditional probabilities [30] while also least- dependent-component analysis is in the focus [55]. New approaches incorporate information theoretic principles directly: Pham investigated BSS based on mutual information [47], whereas Minami applied β-divergences [44]. A method for learning overcomplete data representations and performing overcomplete noisy blind source separation is the sparse coding neural gas (SCNG) [36].

Related to statistical independence, the more difficult problem of estimating statistical dependence is becoming increasingly important and novel algorithms are becoming available [52],[54],[53]. Their applicability is enormous, ranging from variable selection, to BSS to statistical causality.

A comprehensive overview for non-negative matrix and tensor factorization is the book by Cichocki & Amari [16]. Recent results including modern divergences (generalized α-β-divergences were just published [15],[14]. Further, it should be noticed that an information theoretic divergence measure like Rényi-divergences (belonging to the family of α-divergences) capture directly the statistical information contained in the data as expressed by the probability density function.

Otherwise, information theoretic values like the mutual information can be explicitly estimated from data [34]. Here, standard approaches of machine learning such as topographic maps or kernels are applied to achieve accurate estimators [45],[61],[60]. Jenssen et al. have established equivalencies between kernel methods and information theoretic methods [33]. Another example in information theoretic learning uses Rényi-entropy as a cost function instead of the mean squared error, which can be determined either by Parzen estimation [48] or on the basis of the nearest neighbor entropy estimation model [40].

Generally, the question in this context can be translated as: how to deal with the uncertainty contained in data. One way in this direction is to interpret the data as fuzzy values or to generate information about data equipped with uncertainty. This could be done in probabilistic terms, as for example by multivariate

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

class labeling [51], or more general fuzzy approaches.

3 Information theoretic feature extraction and selection

Feature extraction or selection can be seen as a kind of dimension reduction of complex data by constructing combinations of a few variables. These variables should lead to a simplification of the data in a given context but still describing it with sufficient accuracy. By removing most irrelevant and redundant features, feature selection and extraction helps to improve the performance of learning models. It is clear that the complexity of finding an optimal solution grows with the number of features exponentially. Yet, frequently only a sufficient good solution is required. Most feature selection approaches are supervised schemes such that class information or expected regression values can be used for constructing such suboptimal feature subsets or respective ranking list. The strategies to achieve this goal can be classical Bayesian inference schemes [42], or statistical approaches like correlation or covariance investigation [56],[46].

An alternative to these approaches is feature extraction using the non-parametric mutual information. This can be realized by explicit maximization of the respective mutual information [57], or by learning of appropriate feature transformation optimizing the mutual information based on Rényi-entropies [58]. Andonie & Cataron suggested the utilization of a kind of information energy for relevance learning, which is structural similar to the Rényi-entropy but different in detail [4]. Thereby, relevance learning is a more general approach of input variable weighting in learning vector quantization [27]. Introducing sparseness constraints in this scheme according to the Occam's razor principle, the sparsity can be expressed in terms of entropy and, therefore, used for respective optimization [66].

Information theoretic feature selection for *functional data* classification is investigated in [25] based on mutual information optimization while forward-backward strategy search for regression problems based on mutual information is studied in [24].

4 Information theoretic approaches for vector quantization

Vector quantization by a set of prototypes $\mathbf{w}$ is one of the most prominent methods for clustering and data compression based on the optimization of the γ-reconstruction error $E_{VQ}(\gamma)$. One of the key results concerning information theoretic principles for vector quantization is the magnification law discovered by Zador [23],[68]: If the data are given as vectors $\mathbf{v}$ in q-dimensional Euclidean space according to a probability density P and ρ is the probability, then the magnification law $\rho \sim P^{\alpha}$ holds with the magnification factor $\alpha = \frac{q}{q+\gamma}$ related to the reconstruction error according to

$$E_{VQ}(\gamma) = \int \|\mathbf{v} - \mathbf{w}(\mathbf{v})\|_E^{\gamma} P(\mathbf{v}) d(\mathbf{v})$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

with $\|\mathbf{v} - \mathbf{w}(\mathbf{v})\|_E$ is the Euclidean distance of the data vector and the prototype $\mathbf{w}(\mathbf{v})$ representing it. This is the basic principle of vector quantization based on Euclidean distances. For different schemes like self-organizing maps, neural gas variants slightly different magnification factors are obtained due to the neighborhood cooperativeness during prototype adaptation [62],[26],[43]. Optimum magnification is obtained for $\alpha = 1$, which is equivalent to maximum mutual information [68]. Yet, it is possible to control the magnification of most of these algorithm by different strategies such as local learning, winner relaxing or frequency sensitive competitive learning [2],[19]. For a overview we refer to [62].

If divergences are used instead of the Euclidean norm, optimum magnification $\alpha = 1$ can also be achieved by maximum entropy learning [65]. Recently, an approximation to $\alpha = 1$ was also obtained when the mean square error in the self-organizing map training is substituted by correntropy [13]. Divergence measure captures directly the statistical information contained in the data as expressed by the probability density function and can thus produce non-convex cluster boundaries. Generally, vector quantization using different types of divergences as similarity measure is an actual hot topic [5],[31]. A comprehensive overview is given in [64].

Other information theoretic vector quantizers directly optimize the mutual information or strongly related the Kullback-Leibler-divergence. These approaches do not try to minimize the reconstruction error but reduce the divergence between data and prototype density distributions [28] or maximizing unconditional entropy [59]. Vector quantization algorithm directly derived from information theoretic principles based on Rényi-entropies are intensively studied in [38],[49],[22] also highlighting its connection to graph clustering and Mercer kernel based learning [32].

5 Information theory based data visualization and shape recognition

Data visualization is a challenging task to explore data and extract information content. Mapping of complex data structures or high-dimensional data onto the two-dimensional plane or the three-dimensional space preserving the relevant information is of particular interest [**?**]. A good way is standard principal component analysis or its nonlinear counterpart [**?**]. Another option are topographic maps like the above mentioned self-organizing map (SOM) or generative topographic mapping (GTM) [8]. Compared to SOM, also the magnification properties of GTM mapping are known relating them to information preserving mapping [7]. Structural visualization based on SOMs is recently published and denoted as Exploration Machine (XOM) [67], which can be seen as variant of multi-dimensional scaling (MDS) [20]. Both approaches use in original the discrepancy between the pairwise data distances in the data space and in the embedding space based on the Euclidean distances. Yet, also divergences as dissimilarity measure in MDS is proposed [37].

Stochastic neighbor embedding (SNE) provides a principle alternative to

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

MDS: here, the distributions the embedding space is under control such that the mutual information is maximized, which is equivalent to minimizing the Kullback-Leibler-divergence between them [29]. A more robust variant is t-SNE, which uses a t-distribution instead of Gaussians in the original SNE [41]. A mathematical foundations of generalizations of t-SNE and SNE for arbitrary divergences is given in [63]. Yet, Kullback-Leibler-divergence can also be plugged into XOM [10],[11]. A generalization of these ideas leads to self-organized neighbor embedding (SONE) [12],[9]

6 Conclusion

This introduction paper reviews some recent developments in information related learning. Obviously, this paper can not be complete in any sense. However, it highlights some interesting new details and ideas in the field of a rapidly developing field. Information related learning is, thereby, a very general concept, which makes less assumptions about data than many other machine learning approaches and provides principled strategies based on fundamental cognizance about nature. As we have seen, it comprises different methodologies implicitly or explicitly making use of the concepts of information and entropy.

References

[1] D. Ackley, G. Hinton, and T. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–189, 1985.

[2] S. C. Ahalt, A. K. Krishnamurty, P. Chen, and D. E. Melton. Competitive learning algorithms for vector quantization. *Neural Networks*, 3(3):277–290, 1990.

[3] S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Springer, 1985.

[4] R. Andonie and A. Cataron. An information energy LVQ approach for feature ranking. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks 2004*, pages 471–476. d-side publications, 2004.

[5] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[6] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–110, 1943.

[7] C. M. Bishop, M. Svensen, and C. K. I. Williams. Magnification factors for the SOM and GTM algorithms. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6*, pages 333–338. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

[8] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[9] K. Bunte, S. Haase, M. Biehl, and T. Villmann. Mathematical foundations of self-organized neighbor embedding (SONE) for dimension reduction and visualization. *Machine Learning Reports*, 4(MLR-03-2010):1–21, 2010.

[10] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller. Exploratory observation machine (XOM) with kullback-leibler divergence for dimensionality reduction and visualziation. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN'2010)*, pages 87–92, Evere, Belgium, 2010. d-side publications.

[11] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller. Neighbor embedding XOM for dimension reduction and visualization. *Neurocomputing*, page in press, 2011.

[12] K. Bunte, F.-M. Schleif, S. Haase, and T. Villmann. Mathematical foundations of the self organized neighbor embedding (SONE) for dimension reduction and visualization. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN'2011)*, page in this volume, Evere, Belgium, 2011. d-side publications.

[13] R. Chalasani and J. Principe. Self organizing maps with the correntropy induced metric. In *Proc. Int. Joint. Conf. Neural Networks(IJCNN)*, Barcelona, Spain, 2010.

[14] A. Cichocki and S. C. S.-I. Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13:134–170, 2011.

[15] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 2008.

[16] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. Wiley, Chichester, 2009.

[17] G. Deco, W. Finnoff, and H. Zimmermann. Unsupervised mutual information criterion for elemination of overtraining in supervised mulilayer networks. *Neural Computation*, 7:86–107, 1995.

[18] G. Deco and D. Obradovic. *An Information-Theoretic Approach to Neural Computing*. Springer, Heidelberg, New York, Berlin, 1997.

[19] D. DeSieno. Adding a conscience to competitive learning. In *Proc. ICNN'88, International Conference on Neural Networks*, pages 117–124, Piscataway, NJ, 1988. IEEE Service Center.

[20] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

[21] S. Eguchi. Information divergence geometry and the application to statistical machine learning. In F. Emmert-Streib and M. Dehmer, editors, *Information Theory and Statistical Learning*, pages 309–332. Springer S, New York, 2009.

[22] E. Gokcay and J. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–170, 2002.

[23] S. Graf and H. Lushgy. *Foundations of quantization for random vectors*. LNM-1730. Springer, Berlin, 2000.

[24] A. Guillén, A. Sorjamaa, G. Rubio, A. Lendasse, and I. Rojas. Mutual information based initialization of forward-backward search for feature selection in regression problems. In C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, editors, *Proc. of the International Conference on Artificial Neural Networks - ICANN 2009*, volume Part I of *LNCS*, pages 1–9, Berlin / Heidelberg, 2009. Springer.

[25] V. Gómez-Verdejo, M. Verleysen, and J. Fleury. Information-theoretic feature selection for functional data classification. *Neurocomputing*, 72(16-18):3580–3589, 2009.

[26] B. Hammer, A. Hasenfuss, and T. Villmann. Magnification control for batch neural gas. *Neurocomputing*, 70(7-9):1225–1234, March 2007.

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

[28] A. Hegde, D. Erdogmus, T. Lehn-Schioler, Y. Rao, and J. Principe. Vector quantization by density matching in the minimum Kullback-Leibler-divergence sense. In *Proc. of the International Joint Conference on Artificial Neural Networks (IJCNN) - Budapest*, pages 105–109, IEEE Press, 2004.

[29] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840. The MIT Press, Cammbridge, MA, USA, 2002.

[30] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. J. Wiley & Sons, 2001.

[31] E. Jang, C. Fyfe, and H. Ko. Bregman divergences and the self organising map. In C. Fyfe, D. Kim, S.-Y. Lee, and H. Yin, editors, *Intelligent Data Engineering and Automated Learning IDEAL 2008*, LNCS 5326, pages 452–458. Springer, 2008.

[32] R. Jenssen, D. Erdogmus, J. Principe, and T. Eltoft. The Laplacian PDF distance: A cost function for clustering in a kernel feature space. In *Advances in Neural Information Processing Systems*, volume 17, pages 625–632, Cambridge MA, USA., 2005.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

[33] R. Jenssen, D. Erdogmus, J. Principe, and T. Eltoft. Some equivalences between kernel methods and information theoretic methods. *Journal of VLSI Signal Processing*, 45:49–65, 2006.

[34] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):66–138, 2004.

[35] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[36] K. Labusch, E. Barth, and T. Martinetz. Sparse coding neural gas: Learning of overcomplete data representations. *Neuro*, 72(7-9):1547–1555, 2009.

[37] P. Lai and C. Fyfe. Bregman divergences and multi-dimensional scaling. In M. Köppen, N. Kasabov, and G. Coghill, editors, *Proceedings of the International Conference on Information Processing 2008 (ICONIP)*, volume Part II of *LNCS 5507*, pages 935–942. Springer, 2009.

[38] T. Lehn-Schiøler, A. Hegde, D. Erdogmus, and J. Principe. Vector quantization using information theoretic concepts. *Natural Computing*, 4(1):39–51, 2005.

[39] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transaction on Information Theory*, 52(10):4394–4412, 2006.

[40] E. Liitiäinen, A. Lendasse, and F. Corona. On the statistical estimation of rényi entropies. In *Proceedings of IEEE/MLSP 2009 International Workshop on Machine Learning for Signal Processing, Grenoble (France)*, 2009.

[41] L. Maaten and G. Hinten. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[42] D. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[43] E. Merényi, A. Jain, and T. Villmann. Explicit magnification control of self-organizing maps for "forbidden" data. *IEEE Transactions on Neural Networks*, 18(3):786–797, May 2007.

[44] M. Minami and S. Eguchi. Robust blind source separation by beta divergence. *Neural Computation*, 14:1859–1886, 2002.

[45] Y.-I. Moon, B. Rajagopalan, and U. Lall. Estimating mutual inf ormation by kernel density estimators. *Physical Review E*, 52:2318–2321, 1995.

[46] M.Strickert, B. Labitzke, A. Kolb, and T. Villmann. Multispectral image characterization by partial generalized covariance. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN'2011)*, page in this volume, Evere, Belgium, 2011. d-side publications.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

[47] D. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Transactions on Information Theory*, 48:1935–1946, 2002.

[48] J. Principe. *Information Theoretic Learning*. Springer, Heidelberg, 2010.

[49] J. C. Principe, J. F. III, and D. Xu. Information theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering*. Wiley, New York, NY, 2000.

[50] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386–408, 1958.

[51] P. Schneider, T. Geweniger, F.-M. Schleif, M. Biehl, and T. Villmann. Multivariate class labeling in Robust Soft LVQ. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN'2011)*, page in this volume, Evere, Belgium, 2011. d-side publications.

[52] S. Seth, I. Park, and J. Principe. Variable selection: A statistical dependence perspective. In *Proc. International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2010.

[53] S. Seth and J. Principe. A test of Granger non-causality based on nonparametric conditional independence. In *Proc. International Conference on Pattern Recognition (ICPR)*, 2010.

[54] S. Seth and J. Principe. Variable selection: A statistical dependence perspective. In *Proc. International Conference on Machine Learning Applications (ICMLA)*, 2010.

[55] H. Stogbauer, A. Kraskov, S. Astakhov, and P. Grassberger. Least-dependent-component analysis based on mutual information. *Physical Review E*, 70(6):66–123, 2004.

[56] M. Strickert, F.-M. Schleif, U. Seiffert, and T. Villmann. Derivatives of pearson correlation for gradient-based analysis of biomedical data. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, (37):37–44, 2008.

[57] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[58] K. Torkkola and W. Campbell. Mutual information in learning feature transformations. In P. Langley, editor, *Proc. Of International Conference on Machine Learning ICML'2000*, Stanford, CA, 2000. Morgan Kaufmann.

[59] M. M. van Hulle. Topographic map formation by maximizing unconditional entropy: a plausible strategy for 'on-line' unsupervised competitive learning and nonparametric density estimation. *IEEE Transactions on Neural Networks*, 7(5):1299–1305, 1996.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

[60] M. M. van Hulle. Density-based clustering with topographic maps. *IEEE Transactions on Neural Networks*, 10(1):204–207, 1999.

[61] M. M. van Hulle. *Faithful Representations and Topographic Maps From Distortion- to Information-based Self-organization*. J. Wiley & Sons, Inc., 2000.

[62] T. Villmann and J.-C. Claussen. Magnification control in self-organizing maps and neural gas. *Neural Computation*, 18(2):446–469, February 2006.

[63] T. Villmann and S. Haase. Mathematical foundations of the generalization of t-SNE and SNE for arbitrary divergences. *Machine Learning Reports*, 4(MLR-02-2010):1–16, 2010. ISSN:1865-3960, http://www.uni-leipzig.de/~compint/mlr/mlr$_0$2$_2$010.*pdf*.

[64] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, page in press, 2011.

[65] T. Villmann and S. Haase. Magnification in divergence based neural maps. In R. Mikkulainen, editor, *Proceedings of the International Joint Conference on Artificial Neural Networks (IJCNN 2011)*, page in press, San Jose, California, 2011. IEEE Computer Society Press, Los Alamitos.

[66] T. Villmann and M. Kästner. Sparse functional relevance learning in generalized learning vector quantization. In T. Honkela and E. Oja, editors, *Proc. of Workshop on Self-Organizing Maps (WSOM'2011)*, page in press, Helsinki, Finland, 2011. Springer.

[67] A. Wismüller. The exploration machine – a novel method for data visualization. In J. Principe and R. Miikkulainen, editors, *Advances in Self-Organizing Maps – Proceedings of the 7th InternationalWorkshop WSOM 2009, St. Augustine, FL, USA*, LNCS5629, pages 344–352, Berlin, 2009. Springer.

[68] P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transaction on Information Theory*, (28):149–159, 1982.