ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Relevance learning in generative topographic maps

Andrej Gisbrecht and Barbara Hammer

Clausthal University of Technology, Department of Computer Science
D-38678 Clausthal-Zellerfeld, Germany

Abstract. The generative topographic map (GTM) provides a flexible statistical model for unsupervised data inspection and topographic mapping. However, it shares the property of most unsupervised tools that noise in the data cannot be recognized as such and, in consequence, is visualized in the map. The framework of relevance learning or learning metrics as introduced in [4, 6] offers an elegant way to shape the metric according to auxiliary information at hand such that only those aspects are displayed in distance-based approaches which are relevant for a given classification task. Here we introduce the concept of relevance learning into GTM such that the metric is shaped according to auxiliary class labels. Relying on the prototype-based nature of GTM, several efficient realizations of this paradigm are developed and compared on a couple of benchmarks.

1 Introduction

The GTM has been introduced as a generative statistical model corresponding to the classical self-organizing map for unsupervised data inspection and topographic mapping [2]. An explicit statistical model has the benefit of great flexibility and easy adaptability to complex situations by means of statistical assumptions which are fitted to the situation at hand. Like standard unsupervised machine learning and data inspection methods, however, GTM shares the 'garbage in - garbage out' problem: the information inherent in the data is displayed independent of the specific user intention. Hence, if 'garbage' is present in the data, this noise is presented to the user since the statistical model has no way to identify the noise as such.

To partially prevent this fundamental problem of unsupervised data inspection methods, the principle of learning metrics has been introduced into the self-organizing map and alternative data projection schemes in [6]. Thereby, auxiliary information such as class labels are integrated and only those aspects of the data are displayed which carry information for the given auxiliary data at hand. This way, the user can control the aspects which are displayed in the model by providing appropriate information. From a technological point of view, the integration of auxiliary information is realized by means of an adaptation of the metric which determines the map. The Riemannian metric is adapted such that the aspects relevant for the auxiliary information determine the distance computation. Since the computation of a full Riemannian metric leads to rather complex path integrals, the methods as presented in [6, 5] rely on different approximations of the computation. In contrast, metric adaptation in supervised prototype based methods as presented e.g. in [4, 7] introduces global distances or distances attached to the receptive fields of prototypes. This way, a very efficient metric computation takes place and metric parameters can be adapted according to the cost function of the given supervised learning scheme.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

In this contribution, we extend GTM to the principle of learning metrics by combining the technique of relevance learning as introduced in supervised prototype-based classification schemes and the prototype-based unsupervised representation of data as provided by GTM. We propose different ways to adapt the relevance terms which rely on different cost functions connected to prototype-based classification of data. Unlike [3], where a separate supervised model is trained to arrive at appropriate metrics for unsupervised data visualization, we can directly integrate the metric adaptation step into GTM due to the prototype-based nature of GTM. We test the ability of the model to visualize and cluster given data sets on a couple of benchmarks. It turns out that, this way, an efficient and flexible discriminative data mining and visualization technique arises.

2 The generative topographic map

The GTM as introduced in [2] models data $\mathbf{x} \in \mathbb{R}^D$ by means of a mixture of Gaussians which is induced by a lattice of points $\mathbf{w}$ in a low dimensional latent space which can be used for visualization. The lattice points are mapped via a function $\mathbf{w} \mapsto \mathbf{t} = y(\mathbf{w}, \mathbf{W})$ to the data space, where the function is parameterized by $\mathbf{W}$; one can, for example, pick a generalized linear regression model based on base functions such as Gaussians. Every latent point induces a Gaussian

$$p(\mathbf{x}|\mathbf{w}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\|\mathbf{x} - y(\mathbf{w}, \mathbf{W})\|^2\right) \tag{1}$$

with bandwidth β, which give the data distribution as mixture of K modes

$$p(\mathbf{x}|\mathbf{W}, \beta) = \sum_{k=1}^{K} p(\mathbf{w}^k) p(\mathbf{x}|\mathbf{w}^k, \mathbf{W}, \beta)$$

where, usually, $p(\mathbf{w}^k)$ is taken as uniform distribution of the prototypes. Training of GTM optimizes the data log-likelihood

$$\ln\left(\prod_{n=1}^{N}\left(\sum_{k=1}^{K} p(\mathbf{w}^k) p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)\right)\right)$$

by means of an EM approach with respect to the parameters $\mathbf{W}$ and β. In the E step, the responsibility of mixture component k for data point n is determined as

$$r^{kn} = p(\mathbf{w}^k|\mathbf{x}^n, \mathbf{W}, \beta) = \frac{p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta) p(\mathbf{w}^k)}{\sum_{k'} p(\mathbf{x}^n|\mathbf{w}^{k'}, \mathbf{W}, \beta) p(\mathbf{w}^{k'})} \tag{2}$$

while an algebraic expression for $\mathbf{W}$ and β can be derived in the M step [2].

3 Relevance learning

The principle of relevance learning has been introduced in [4] as a particularly simple and efficient method to adapt the metric of prototype based classifiers

INIT
REPEAT
  E-STEP: DETERMINE THE RESPONSIBILITIES r^{kn} BASED ON $\|\mathbf{x}-\mathbf{w}\|^2_\lambda$
  M-STEP: DETERMINE W AND β AS IN GTM
  LABEL PROTOTYPES
  ADAPT λ BY STOCHASTIC GRADIENT DESCENT ON $E(\lambda)$
  NORMALIZE λ

Table 1: Integration of relevance learning into GTM

according to the given situation at hand. It takes into account a relevance scheme of the data dimensionalities by substituting the euclidean metric by the weighted form

$$\|\mathbf{x}-\mathbf{w}\|^2_\lambda = \sum_{d=1}^{D} \lambda_d^2 (x_d - w_d)^2 \,. \tag{3}$$

In [4], the euclidean metric is substituted by the more general form (3) and, parallel to prototype updates, the metric parameters λ are adapted according to the given classification task. Here, we introduce the same principle into GTM.

Assume that data point $\mathbf{x}$ is equipped with label information y which is element of a finite set of different labels. Prototypes of a given GTM can be labeled posteriorly based on this information, i.e. prototype $\mathbf{t}^k = y(\mathbf{w}^k, \mathbf{W})$ is labelled

$$c(\mathbf{t}^k) = \arg\max_c \left(\sum_{n|y^n=c} r^{kn} \right) \tag{4}$$

We can introduce relevance learning into GTM by substituting the euclidean metric in the Gaussian bells (1) by the more general diagonal metric (3) which includes relevance terms. Analogous to [2], it can be seen that optimization of the parameters $\mathbf{W}$ and β of GTM can be done the same way as beforehand, whereby the diagonal metric has to be used when computing the responsibilities (2). The metric parameters λ should be adapted such that the label information is taken into account. Below, we will introduce different objective functions which are motivated by the classification induced by the prototype based GTM with posterior labeling (4). Note that, further, it is advisable to normalize the parameters $\|\lambda\|^2 = 1$ to prevent degeneration to the trivial solution $\lambda = 0$. The principled integration of relevance learning into GTM is depicted in Tab. 1. Thereby, usually one epoch is performed to adapt the relevance terms by means of a stochastic gradient descent of an appropriate cost function $E(\lambda)$. Now, we discuss concrete cost functions $E(\lambda)$ for the relevance terms in more detail.

Generalized Relevance GTM (GRGTM)

The cost function is taken from generalized relevance learning vector quantization, which can be interpreted as the goal to maximize the hypothesis margin of a prototype based classification scheme such as LVQ [4, 7]. The cost function is given as

data	number of prototypes	number of base functions
Landsat	10×10	14×14
Phoneme	10×10	5×5
Letter	30×30	30×30

Table 2: Parameters used for training

$$E(\lambda) = \sum_n \mathrm{sgd}\left(\frac{\|\mathbf{x}^n - \mathbf{t}^+\|^2_\lambda - \|\mathbf{x}^n - \mathbf{t}^-\|^2_\lambda}{\|\mathbf{x}^n - \mathbf{t}^+\|^2_\lambda + \|\mathbf{x}^n - \mathbf{t}^-\|^2_\lambda}\right)$$

where $\mathbf{t}^+$ is the closest prototype in the data space with the same label as $\mathbf{x}^n$ and $\mathbf{t}^-$ is the closest prototype with a different label.

Robust Soft GTM (RSGTM)

In analogy to soft robust LVQ as introduced in [8], the goal is to optimize the ratio of the probability of correct classification to the overall probability:

$$E(\lambda) = \sum_n \log\left(\frac{\sum_{k|c(\mathbf{t}^k)=y^n} p(\mathbf{w^k})p(\mathbf{x}^n|\mathbf{w^k}, \mathbf{W}, \beta)}{p(\mathbf{x}^n|\mathbf{W}, \beta)}\right)$$

Entropy GTM (EGTM)

The goal is to obtain prototypes such that labels of points assigned to these prototypes coincide as much as possible. This can be measured by means of the entropy of the label distribution assigned to the prototypes:

$$E(\lambda) = -\sum_k \sum_c \frac{\sum_{n|y^n=c} p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)}{\sum_n p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)} \log \frac{\sum_{n|y^n=c} p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)}{\sum_n p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)}$$

In all three cases, update rules can be derived from $E(\lambda)$ by taking the derivative with respect to λ. This way, we obtain three different update schemes for the relevance terms, which are based on different fundamental principles connected to the classification induced by GTM: margin maximization, optimization of the probability ratio of correct classification, and optimization of the class entropy attached to the prototypes, respectively.

4 Experiments

We test the efficiency of relevance learning in GTM on three benchmark data sets as described in [6]: Landsat Satellite data with 36 dimensions, 6 classes, and 6435 samples, Letter Recognition data with 16 dimensions, 26 classes, and 20000 samples, and Phoneme data with 20 dimensions, 13 classes, and 3656 samples. GTM is initialized using the first two principal components. The mapping $y(\mathbf{w}, \mathbf{W})$ is induced by generalized linear regression based on Gaussian base functions. The learning rate of the gradient descent for λ has been optimized for the data and is chosen in the range of 10^{-6} to 10^{-2}. The number of epochs is

	GTM	GRGTM	SRGTM	EGTM	Graph
Landsat					
accuracy	85.92 (0.2)	86.33 (0.15)	86.18 (0.23)	85.14 (0.46)	88.95
top.product	-0.0087	-0.0080	-0.0101	0.0009	
random	-0.0355	-0.0589	-0.0653	-0.0174	
Phoneme					
accuracy	40.15 (0.32)	76.69 (1.29)	76.62 (2.24)	83.44 (1.03)	90.77
top.product	-0.0391	-0.0098	-0.0365	-0.0131	
random	-0.0953	-0.0307	-0.1093	-0.0783	
Letter					
accuracy	67.15	77.92	78.45	66.23	59.26
top.product	-0.0658	-0.0684	-0.0772	-0.0693	
random	-0.1585	-0.1616	-0.1629	-0.1592	

Table 3: Results obtained with GTM and relevance learning, the standard deviation for the accuracy is displayed in parenthesis. The topographic product is evaluated on the trained map and a random permutation, for comparison. Since random permutation disrupts the topological ordering, the number should be a magnitude larger in size than the evaluation of the topographic product on a topologically sorted map. The column 'Graph' refers to the result as reported in [6] for Sammon's map with full graph-based Riemannian metric adapted to the given labeling.

chosen as 100. The number of prototypes and base functions has been optimized on the data and is shown in Tab. 2. We report the results of a repeated ten-fold cross-validation, whereby we evaluate the models by means of the classification error, to estimate the capability of the models of taking auxiliary label information into account, and by means of the topographic product as introduced in [1] to judge the capability of the models of faithful topological representation of the data. Since the topographic product is sensitive to the number of prototypes used for the models, we always report the average topographic product of randomly perturbed maps in comparison; values close to 0 indicate a faithful topographic ordering of the map. The results obtained this way are depicted in Tab. 3. For comparison, we report the results obtained by GTM without relevance learning, and the classification accuracy which can be obtained by the more demanding principle of (full graph based) learning metrics in combination with Sammon's mapping as reported in [6].

In all cases but one the integration of label information in terms of relevance learning greatly improves the classification accuracy of GTM, while not deteriorating the visualization quality of the map as measured by the topographic product. Interestingly, although the method is restricted to a very simple diagonal metric, the accuracy is comparable to the accuracy obtained by a full Riemannian metric as investigated in [6]. Apart from an improved efficiency (the algorithm being $\mathcal{O}(N)$ only, N denoting the number of data points), global relevance terms have the benefit that the result can be interpreted directly, the size of the components of λ corresponding to the relevance of the respective dimensions, see Fig. 1 for the relevance profiles as obtained for the Letter data set. Interestingly, the methods widely agree on the fact that the dimensions around 8 are of particular importance for this task, although relevance adaptation relies

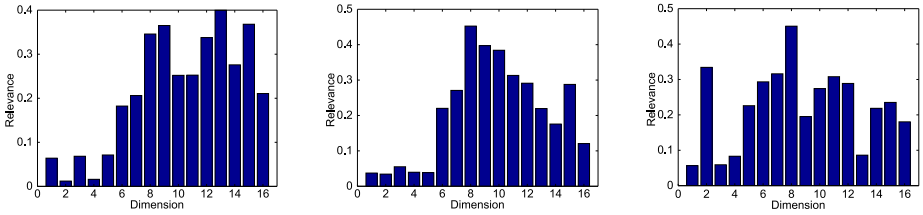

Fig. 1: Relevance profile of GRGTM (left), RSGTM (middle), EGTM (right).

on three different principles for these adaptation rules.

5 Discussion

In this contribution, a method has been proposed to integrate auxiliary information in terms of relevance updates into GTM; the benefit of this approach has been demonstrated on three benchmarks. As [6], the work is based on adaptive metrics to incorporate auxiliary information into the model. Unlike [6], however, the proposed method relies on the prototype-based nature of GTM and transfers the relevance update scheme of supervised learning schemes such as [4, 7] to this setting, resulting in an efficient and interpretable discriminative topological mapping. Obviously, the method could be further extended to even more flexible metrics such as individual matrices attached to the prototypes, as proposed in the frame of supervised learning in [7]. The investigation of these possibilities will be the subject of future work.

References

[1] H.-U. Bauer and K. Pawelzik. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks* 3(4):570-579, 1992.

[2] C. Bishop, M. Svensen, and C. Williams. The generative topographic map. *Neural Computation* 10(1):215-234, 1998.

[3] K. Bunte, B. Hammer, P. Schneider, M. Biehl. Nonlinear Discriminative Data Visualization. *ESANN 2009*, M. Verleysen (ed.), d-side publishing, 65-70, 2009.

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

[5] J. Peltonen. *Data exploration with learning metrics*. D.Sc. thesis. Dissertations in Computer and Information Science, Report D7. Espoo, Finland, 2004.

[6] J. Peltonen, A. Klami, and S. Kaski. Improved Learning of Riemannian Metrics for Exploratory Analysis. *Neural Networks* 17: 1087-1100, 2004.

[7] P. Schneider, M. Biehl, and B. Hammer. Adaptive Relevance Matrices in Learning Vector Quantization. *Neural Computation* 21: 3532-3561, 2009.

[8] S. Seo and K. Obermayer. Soft Learning Vector Quantization. *Neural Computation* 15(7): 1589-1604, 2003.