Relevance learning for short high-dimensional time series in the life sciences

F.-M. Schleif, A. Gisbrecht, and B. Hammer,
University of Bielefeld,
CITEC Center of Excellence,
Universitätsstrasse 21-23,
33615 Bielefeld, Germany
Email: {fschleif|agisbrec|bhammer}@techfak.uni-bielefeld.de

***Abstract*—Digital data characterizing physiological processes over time are becoming increasingly important such as spectrometric data or gene expression profiles. Typical characteristics of such data are high dimensionality due to a fine grained measurement, but usually only few time points of the series. Due to the short length, classical time series models cannot be used. At the same time, due to the high dimensionality, data cannot be treated by means of time windows using simple vectorial techniques.**

Here, we consider the generative topographic mapping through time (GTM-TT) as a highly regularized model for time series inspection in the unsupervised setting, based on hidden Markov models enhanced with topographic mapping facilities. We extend the model such that supervised classification can be built on top of GTM-TT, resulting in supervised GTM-TT, and we extend the technique by supervised relevance learning. The latter adapts the metric according to given auxiliary information resulting in an interpretable form which can deal with high dimensional inputs. We demonstrate the technique in simulated data as well as an example from the biomedical domain, reaching state of the art classification accuracy in both cases.

I. Introduction

More and more data sets occurring in the biomedical domain are electronically available. Improved measurement technology often leads to very high dimensional signals representing a fine grained resolution of the available information, such as e.g. mass spectrometric measurements or gene expressions. At the same time, an increasing amount of data comes with a temporal axis representing several measurements taken over time. Typical examples include measurements over time derived from serum or blood samples to judge the effectiveness of a therapy. In such cases, the measurement at one point in time represents a mass spectrum or a gene expression profile at a dedicated point in time, and the overall information (the effectiveness of a therapy) is mirrored by the development over these measures over a short time span. Due to the high costs of such measurements, typically, less than 10 time points are encountered, while data dimensionality at one time point can easily excess several hundred or thousand dimensions.

The analysis of high-dimensional short time series is a challenging task. Due to the time characteristics, data are no longer identical and independently distributed, and a treatment as independent vectors over time is not appropriate. Due to the very high data dimensionality, it is not feasible to address the data simply as one large vector where all time points are collected. Due to the shortness of the times series, standard time series methods like auto-regressive moving average (ARMA) or extensions thereof (see e.g. [1]) are in general not applicable. Thus, this setting constitutes a challenge to modern data processing.

A few methods have been proposed to model these type of data. In [2] an unsupervised projection technique has been proposed employing a so called temporal context. Temporal data are processed by a Self Organizing Map (SOM) [3] but training is modified such that it depends on the current temporal context. A further unsupervised model has been proposed in [4] using the Generative Topographic Mapping Through Time (GTM-TT) ([5]). Here, hidden variables are introduced to account for the relevance of the different feature dimensions, to account, in a non-discriminative manner, for the explained variance in the data over time. A supervised two-class method solely based on hidden Markov models has been proposed in [6]. It models the two different data distributions by independent HMMs and evaluates the generated models to obtain a ranking of the input dimensions. Subsequently the prediction is improved by selecting a set of features using a wrapper strategy. In [7] a similar approach has been proposed in a semi-supervised scenario, introducing class-wise constraints in the hidden Markov model. The importance of the individual features is determined based on a complex post processing procedure. Another supervised method using all features, based on a Support Vector Machine (SVM) and a Kalman filter, has been proposed in [8].

While the first two approaches turn out to be very effective for unsupervised analysis, the last methods focus on supervised and semi-supervised analysis. The results in [6] are very promising, with 85% prediction accuracy on a real life multiple sclerosis data (MS) set, but they rely on strong assumptions about the underlying HMM structure. The approach in [8] improves this result by $2-5\%$ but it results in a black box scenario, without additional feature selection or the possibility to easily inspect the results. The approach in [7] is evaluated also with respect to the results of [6] achieving improved performance for the MS data set. There is further ongoing work in this field, reflecting the high demand for effective methods dealing with these type of data. The application field is not limited to the bio-medical domain as considered in [6], [7], [9] but covers a broader field of applications also in industry and geo-science as reflected in [4], [2].

Since it is based on Gaussians, GTM crucially depends on the underlying metric in observation space, usually the Euclidean metric. For very high dimensional data, however, Euclidean distances become more and more meaningless due to accumulated noise [10]. Further, the relevance of the measured dimensions for the observed task is not clear a priori, and the overall quality of the method severely depends on the fact that the relevance of the dimensions corresponds to their contribution to the metric. Because of this observation, many distance based learning techniques have been extended to automatic relevance adaptation which automatically adapts metric parameters according to given auxiliary information, see e.g. [11], [12].

In this contribution, we are interested in the question how relevance learning can be transferred to the temporal domain. The identification of relevant input dimensions of a temporal sequence is very important as outlined e.g. in [4], [6] to obtain a better understanding of the data, to reduce the processing complexity, and to improve the overall prediction accuracy. As already motivated by some of the prior references, prototype methods (see e.g. [3]) have been found to be very effective for the analysis of high dimensional data also to analyze temporal sequences. In [5], the *Generative Topographic Mapping - through time* (GTM-TT), an unsupervised prototype based method for the topographic projection of high-dimensional, temporal sequences, has been proposed. GTM-TT learns a hidden Markov model (HMM) of a data generating process and represents the data by a prototype based representation in time and space. Like in ordinary prototype methods the GTM-TT approximates the data distribution by a vector quantization of the data space. The temporal dependency between the prototype is modeled by an internal HMM on the latent space. Additionally the prototypes are assigned to a fixed grid representation or lattice, which permits, provided the topology is preserved (see [13]), the easy visualization and interpretation of the data trajectory in a low dimensional space.

In this contribution we extend the GTM-TT to a supervised method and we integrate relevance learning based on the auxiliary class labels. The latter techniques greatly enhances the quality of the map by means of a supervised adaptation of the underlying metric with respect to large margin classification. The resulting metric form is diagonal, i.e. metric parameters can directly be interpreted as relevance of the corresponding input dimension. We demonstrate the effectiveness of the technique in an artificial scenario as well as a real life data set (the MS data set), reaching state-of-the-art classification accuracy in the latter case. In addition, the relevance learning for supervised GTM-TT allows to visualize the classes over time and to directly inspect the resulting relevance profile. Now, we first review GTM and GTM-TT and we explain how to extend this method to supervised scenarios. Then, we explain the principle of relevance learning and transfer it to our setting. Finally, the approach is tested in experiments, discussing open issues and further potential in the conclusions.

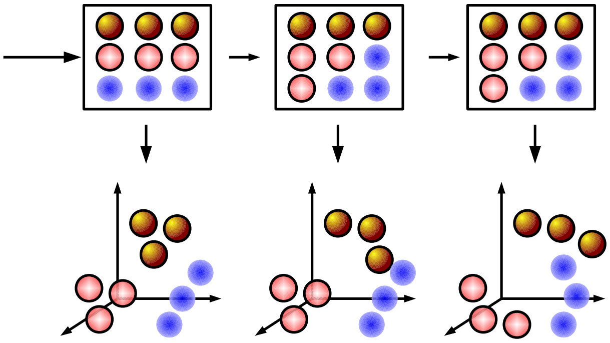


Fig. 1. GTM-TT consisting of a HMM in which the hidden states are given by the latent points of the GTM model. The emission probabilities are governed by the GTM mixture distribution [5]. The different data distributions, exemplified in 3D (bottom) and indicated by the color/shading are mapped to the 2D grid (top). Here we consider 9 hidden states on a 3×3 grid. The data distribution may change over time and hence also the mapping of the GTM is effected over time, assuming smooth transitions.

II. Generative Topographic Mapping

The Generative Topographic Mapping (GTM) as introduced in [14] 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 $\mathbf{w} \mapsto \mathbf{t} = y(\mathbf{w}, \mathbf{W})$ to the data space, where the function is parametrized by $\mathbf{W}$; usually, a generalized linear regression model is chosen

$$y : \mathbf{w} \mapsto \Phi(\mathbf{w}) \cdot \mathbf{W} \tag{1}$$

where the base functions Φ are often chosen as equally spaced Gaussians.The high-dimensional points $y(\mathbf{w}, \mathbf{W})$ are so called prototypes of the original data space, representing the data space as accurately as possible. They can be directly inspected and permit to summarize the data.

Every latent point induces a Gaussian

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

with variance β^{-1}. Assuming a Dirac distribution of the prototypes, this gives the data distribution as mixture of K modes

$$p(\mathbf{x}|\mathbf{W}, \beta) = \sum_{k=1}^{K} p(\mathbf{w}^k) p(\mathbf{x}|\mathbf{w}^k, \mathbf{W}, \beta) \tag{3}$$

with $p(\mathbf{w}^k) = 1/K$. Training of GTM optimizes the data log-likelihood

$$\ln\left(\prod_{n=1}^{N}\left(\sum_{k=1}^{K} p(\mathbf{w}^k) p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)\right)\right) \tag{4}$$

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

$$r^{kn} = p(\mathbf{w}^k|\mathbf{x}^n, \mathbf{W}, \beta) = \frac{p(\mathbf{x}^n|\mathbf{w}^k, \mathbf{W}, \beta)p(\mathbf{w}^k)}{\sum_{k'} p(\mathbf{x}^n|\mathbf{w}^{k'}, \mathbf{W}, \beta)p(\mathbf{w}^{k'})} \tag{5}$$

In the M step, the weights $\mathbf{W}$ are determined solving the equality

$$\mathbf{\Phi}^T \mathbf{G}_{\text{old}} \mathbf{\Phi} \mathbf{W}^T_{\text{new}} = \mathbf{\Phi}^T \mathbf{R}_{\text{old}} \mathbf{X} \tag{6}$$

where $\mathbf{\Phi}$ refers to the matrix of base functions Φ evaluated at points $\mathbf{w}^k$, $\mathbf{X}$ to the data points, $\mathbf{R}$ to the responsibilities, and $\mathbf{G}$ is a diagonal matrix with accumulated responsibilities $G_{nn} = \sum_k r^{kn}(\mathbf{W}, \beta)$. The variance results from

$$\frac{1}{\beta_{\text{new}}} = \frac{1}{ND} \sum_{k,n} r^{kn}(\mathbf{W}_{\text{old}}, \beta_{\text{old}}) \|\Phi(\mathbf{w}^k)\mathbf{W}_{\text{new}} - \mathbf{x}^n\|^2 \tag{7}$$

where data dimensionality D and number of points N.

This way, an unsupervised restricted Gaussian mixture model induced by a low dimensional latent space results. The degree of topology preservation is determined by the stiffness of the mapping y. It can be directly controlled by the number of base functions.

III. GTM Through-Time

The GTM through time (GTM-TT) has been introduced in [5] to extend the topographic mapping to time series data for which the entries are no longer independent. We assume that data are time series in the D-dimensional Euclidean space, i.e. $\mathbf{x} = \mathbf{x}(1) \dots \mathbf{x}(T) \in (\mathbb{R}^D)^*$ where $T \geq 1$ is the length of the time series. A data point of the training set is referred to by $\mathbf{x}^i$. For such data consecutive entries $\mathbf{x}(t)$ and $\mathbf{x}(t+1)$ are strongly correlated. While the data space of observations over time is represented by a topographic mapping as before, this temporal dependency is learned from the data in form of a hidden Markov model. The hidden states are thereby characterized by the lattice points $\mathbf{w}^j$ in the latent space.

The structure of the GTM-TT is shown in Figure 1. Assuming a sequence $\mathbf{x}$ of observations and a sequence of hidden states of the same length $\mathbf{z} = \mathbf{z}(1) \dots \mathbf{z}(T)$ where $\mathbf{z}(i)$ equals a point $\mathbf{w}^j$, the probability of the observations and a specific path of hidden states $\mathbf{z}$ is given by $p(\mathbf{x}, \mathbf{z}|\Theta) =$

$$p(\mathbf{z}(1)) \prod_{t=2}^{T} p(\mathbf{z}(t)|\mathbf{z}(t-1), \mathbf{W}, \beta) \prod_{t=1}^{T} p(\mathbf{x}(t)|\mathbf{z}(t)) \tag{8}$$

where the conditional probability

$$p(\mathbf{x}(t)|\mathbf{z}(t)) := p(\mathbf{x}(t)|\mathbf{z}(t), \mathbf{W}, \beta) \tag{9}$$

is as before (2) [14], resulting in the overall probability of $\mathbf{x}$:

$$p(\mathbf{x}|\Theta) = \sum_{\mathbf{z} \in \{\mathbf{w}^1, \dots, \mathbf{w}^K\}^T} p(\mathbf{x}, \mathbf{z}|\Theta) \tag{10}$$

The parameters of GTM-TT are

$$\Theta = (\mathbf{W}, \beta, \pi, \mathbf{P}) \tag{11}$$

with additional parameters for the initial state probabilities

$$\pi = (\pi_j)_{j=1}^K \text{ where } \pi_j = p(\mathbf{z}(1) = \mathbf{w}^j) \tag{12}$$

and transition probabilities

$$\mathbf{P} = (p_{ij})_{i,j=1}^K \text{ where } p_{ij} = p(\mathbf{z}(t) = \mathbf{w}^j|\mathbf{z}(t-1) = \mathbf{w}^i) \tag{13}$$

which characterize the temporal correlations of subsequent states in latent space, relying on the assumption of the standard Markov property and stationarity of the dynamics.

Based on these definitions, it is possible to optimize the data log likelihood

$$\ln \left(\prod_{n=1}^{N} p(\mathbf{x}^n|\Theta) \right) \tag{14}$$

with respect to the model parameters Θ by an EM-approach as before. As for HMMs, a forward-backward procedure allows to determine the hidden parameters, the responsibilities of states for a given sequence, in an efficient way [15], based on which the parameters $\mathbf{W}$ and β can be determined as before. We obtain the probability of being in state $\mathbf{w}^k$ at time t, given the observation sequence $\mathbf{x}^n$:

$$r^{kn}(t) = p(\mathbf{z}(t) = \mathbf{w}^k|\mathbf{x}^n, \Theta) = \frac{A_{kt}B_{kt}}{p(\mathbf{x}^n|\Theta)} \tag{15}$$

The forward variable A_{kt} is the joint probability $p(\mathbf{x}^n(1) \dots \mathbf{x}^n(t), \mathbf{z}(t) = \mathbf{w}^k|\Theta)$ given by the following recursive equation:

$$A_{kt} = \sum_{i=1}^{K} A_{it-1} p_{ik} p(\mathbf{x}^n(t)|\mathbf{w}^k, \Theta) \tag{16}$$

and $A_{k1} = \pi_k p(\mathbf{x}^n(1)|\mathbf{w}^k, \Theta)$. The backward variable B_{kt} is the joint probability $p(\mathbf{x}^n(t+1) \dots \mathbf{x}^n(t_n), \mathbf{z}(t) = \mathbf{w}^k|\Theta)$ which can be calculated using the following recursive equation

$$B_{kt} = \sum_{i=1}^{K} p_{ik} p(\mathbf{x}^n(t+1)|\mathbf{w}^i, \Theta) B_{it+1} \tag{17}$$

where $B_{kT} = 1$. The transition parameters of the Markov model can be trained using the standard Baum-Welch training. Details can be found in [16].

For an input time series $\mathbf{x}^n(1) \dots \mathbf{x}^n(T)$, GTM-TT gives rise to a time series of responsibilities $r^{kn}(1) \dots r^{kn}(T)$ of neuron k. Based on these responsibilities, a winner can be determined for every time step t as neuron $\text{argmax}_k r^{kn}(t)$.

IV. Supervised GTM-TT

Assume that the observed time series $\mathbf{x}$ is equipped with label information l which is element of a finite set of different labels $1, \dots L$. We extend GTM-TT to a supervised classification scheme in the following way: given a training set, we train separate GTM-TT for every class, whereby the models are coupled by choosing the same bandwidth β and the same underlying topological structure in the latent space, i.e. the same base functions Φ and the same Dirac distribution on the latent space. The prototype parameters $\mathbf{W}_l$

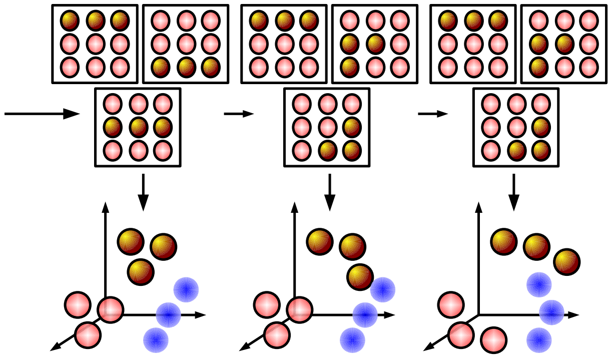


Fig. 2. Illustration of the SGTM-TT. It consists of multiple GTM-TT models. It behaves similar to the regular GTM-TT but the training is classwise and the β parameter is common over the different models. The different classwise models (top) are used to represent the data distribution (bottom) over time (from left to right).

are trained individually for every model representing label l. The same holds for the initial state probability π_l and the transition probabilities $\mathbf{P}_l$. We denote the obtained model as Supervised GTM-TT. The concept of the SGTM-TT is depicted schematically in Figure 2.

When processing a novel time series $\mathbf{x}$ we obtain L time series of responsibilities according to every model. We denote the responsibilities of model l for input $\mathbf{x}$ at time point t by $r_l^k(\mathbf{x}(t))$. This gives rise to an aggregated value of responsibilities for every input series $\mathbf{x}$ and label l:

$$r_l(\mathbf{x}) := \sum_{k=1}^{K} \sum_{t=1}^{T} r_l^k(\mathbf{x}(t))/(KT) \tag{18}$$

One can pick the label l as output for which this value is largest. However, this does not take prior class probabilities into account. Because of this fact, we use an additional linear classifier with inputs given by the vectors $(r_l(\mathbf{x}))_{l=1}^{L}$ which is trained based on the given training set. We define a probabilistic kernel following [17] (p. 297) and train a standard SVM model, involving no further parameters. This way, class priors with maximum discriminative accuracy on the training set are obtained.

V. Relevance learning for SGTM -TT

The principle of relevance learning has been introduced in [11] as a particularly simple and efficient method to adapt the underlying metric of prototype based classifiers according to the given situation at hand. It takes into account a relevance scheme of the data dimensions by substituting the squared Euclidean metric by the weighted form

$$d_{\boldsymbol{\lambda}}(\mathbf{x}, \mathbf{t}) = \sum_{d=1}^{D} \lambda_d^2 (x_d - t_d)^2 \,. \tag{19}$$

The principle is extended in [18], [19] to the more general metric form

$$d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{t}) = (\mathbf{x} - \mathbf{t})^T \mathbf{\Omega}^T \mathbf{\Omega} (\mathbf{x} - \mathbf{t}) \tag{20}$$

Using a square matrix $\mathbf{\Omega}$, a positive semi-definite matrix which gives rise to a valid pseudo-metric is achieved this way. In [18], [19], these metrics are considered in local and global form, i.e. the adaptive metric parameters can be identical for the full model, or they can be attached to every prototype present in the model. Relevance learning for GTM has been already introduced in [12] for i.i.d. data resulting in relevance GTM (R-GTM). In case of temporal sequences some modification of the original principle are necessary and also the supervision will be handled differently as pointed out subsequently.

For simplicity and to maintain easy interpretability, we will restrict to a global diagonal weighting scheme, in which case a weight λ_i directly corresponds to the relevance of dimension i assuming an equal range of the data dimensions. For GTM, the distance used to compute local probabilities is substituted by a weighted distance measure:

$$p_{\lambda}(\mathbf{x}|\mathbf{w}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2} d_{\lambda}(\mathbf{x}, y(\mathbf{w}, \mathbf{W}))\right) \tag{21}$$

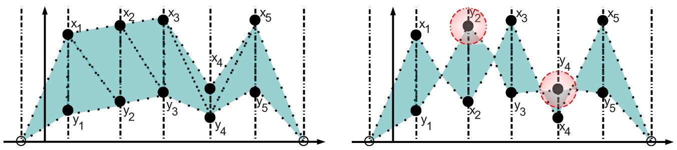
with Euclidean distance which includes relevance terms to weight the single dimensions. This gives rise to a data log likelihood which takes into account the dimensions according to their relevance and, hence, a topographic mapping which mirrors this weighting scheme of the metric.

The question is how to set relevance parameters λ in a way such that the classification accuracy of the resulting mapping is as high as possible. We proceed as in [12] and train the relevance parameters based on priorly given class information in a separate update step which is interleaved with the standard adaptation of the SGTM-TT assuming this changed distance. For relevance learning, we rely on the cost function as introduced in generalized learning vector quantization since it treats the classification as a large margin technique [19]. Metric parameters are adapted such that this margin is optimized.

First, we shortly recall the cost function of GRLVQ for the simple vectorial setting: Assume a prototype based classification is given. Hence a finite set of prototypes $\mathbf{t}^j$ with class labels are present in the data space. Classification takes place by means of a winner takes all scheme, i.e. the label corresponding to the prototype with smallest distance $d_\lambda(\mathbf{x}, \mathbf{t}^j)$ serves as output. For standard GTM, prototypes are induced by latent points $\mathbf{t}^j = y(\mathbf{w}^j, \mathbf{W})$, and the distances determine the responsibilities of the data points. Now adaptation of the relevance terms λ takes place such that the costs

$$E(\lambda) = \sum_{n} \text{sgd}\left(\frac{d_\lambda(\mathbf{x}^n, \mathbf{t}^+) - d_\lambda(\mathbf{x}^n, \mathbf{t}^-)}{d_\lambda(\mathbf{x}^n, \mathbf{t}^+) + d_\lambda(\mathbf{x}^n, \mathbf{t}^-)}\right) \tag{22}$$

are optimized, where $\mathbf{t}^+$ corresponds to the closest prototype with a correct label, whereas $\mathbf{t}^-$ corresponds to the closest prototype with an incorrect label, given input $\mathbf{x}^n$. It has been shown in [19] that this scheme optimizes the hypothesis margin of such a prototype based classifier. It can directly be integrated into vectorial GTM, performing simultaneous adaptation of the parameters of GTM to optimize the data



(a) Two functions: Euc = L^p-norm (b) Two functions: Euc $\neq L^p$-norm

Fig. 3. Illustration of the L^p-norm. Plot (a) indicates the case in which the distance between two functions is equal, both for Euclidean or L^p-norm. In plot (b) parts of the functions are interchanging (crossing). The distance using Euc is still the same as in plot (a) but for the L^p-norm the distance is changed, giving a more realistic measure of the distance of the two functions.

log-likelihood, and the metric parameters to optimize the classification margin. An excellent classification accuracy results as demonstrated in [12]. The adaptation formulas for the parameters λ can be derived from the above cost function by taking the derivatives. To avoid divergence, metric parameters are normalized after every adaptation step.

For SGTM-TT, a problem occurs: the classification is not determined by a single prototype, rather a winner is determined for every class label and every time step t. Classification takes place based on the aggregated responsibilities $r_l(\mathbf{x})$. We use this classification to determine which class is assigned to a given time series. For the metric update, however, we rely on a prototypical representation of time series by SGTM-TT in the following way: for every class label we consider the time series of prototypes of the corresponding GTM-TT model according to the winner prototypes over time, given input sequence $\mathbf{x}$:

$$\mathbf{t}_l = (\mathbf{t}_l(1) \dots \mathbf{t}_l(T)) \tag{23}$$

where

$$\mathbf{t}_l(t) = y(\mathbf{w}^k, \mathbf{W}_l) \text{ with } k = \text{argmax}_k r_l^k(\mathbf{x}(t)) \tag{24}$$

Now we can insert an input time series $\mathbf{x}$ and the corresponding time series of prototypes representing a correct and a wrong class label into the cost function of GRLVQ (22). Assuming an appropriate metric of these two time series, a well defined cost function results. How do we compare time series? As simple way is by averaging over the Euclidean distances in every time point. However, this simple approach completely neglects the functional form of time series. In our experiments, it turned out that a dissimilarity measure which takes into account the functional form as proposed by Lee and Verleysen in [20] is beneficial. It has already successfully been used for the analysis of biomedical data in [21].

The distance measure integrates the functional form of three subsequent time steps to compare the values $\mathbf{x}(t)$ and $\mathbf{t}(t)$ at a given point in time t. Assume there is given a real valued time series $\mathbf{v} = v(1) \dots v(T)$, the functional L_p norm is given by

$$\mathcal{L}_p^f(\mathbf{v}) = \left(\sum_{t=1}^{T} \left(\triangle A_t(\mathbf{v}) + \triangle B_t(\mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{25}$$

with

$$\triangle A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2} |(t)| & \text{if } 0 \leq v(t)v(t-1) \\ \frac{\tau}{2} \frac{(t)^2}{|v(t)| + |v(t-1)|} & \text{if } 0 > v(t)v(t-1) \end{cases} \tag{26}$$

$$\triangle B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2} |v(t)| & \text{if } 0 \leq v(t)v(t+1) \\ \frac{\tau}{2} \frac{v(t)^2}{|v(t)| + |v(t+1)|} & \text{if } 0 > v(t)v(t+1) \end{cases} \tag{27}$$

representing the triangles on the left and right sides of $\mathbf{v}(t)$. Values at the boundaries corresponding to time step 0 or $T+1$ are set to 0. This norm takes into account whether the entries change the sign in subsequent time steps. For vectorial data $\mathbf{x}$ and $\mathbf{t}$ over time with dimensionality D in every time step, this induces the weighted distance

$$d_\lambda(\mathbf{x}, \mathbf{t}) = \sum_{i=1}^{D} \lambda_i \mathcal{L}_p^f(\mathbf{x}_i - \mathbf{t}_i) \tag{28}$$

where $\mathbf{x}_i - \mathbf{t}_i$ refers to the time series of real numbers given by the distance of the entries in dimension i. This distance measure does not only take the absolute distance into account, but also measures whether the curves have the same shape. As before, the relevance of dimension i is weighted by a relevance term λ_i. The concept of the L^p-norm is shown in Figure 3.

The weighted distance (28) is now inserted into the cost function of GRLVQ. Taking the derivatives (see [20] for the derivatives of the L_p norm) with respect to the relevance terms yields an adaptive weighting scheme for the input dimensions which takes the functional form of the data into account. As before, λ is normalized after every adaptation step to guarantee nonnegative values which sum up to 1.

VI. Experiments

Subsequently we consider two data sets to evaluate our approach.

A. Data and parameter settings

1) Simulated data set: The first one is a simulated two class scenario, proposed in the paper [6]. It consists of 100 samples divided into two classes of 50 samples each. For each sample 100 features have been generated with 8 time points. Out of the 100 features, only 10 where substantially differentiating between the classes. The generation mechanism behind the simulated data is to sample the time series from a piecewise linear function. At a later step, sample-specific variation is included by shrinking and expanding the curves.

2) Multiple sclerosis data: The second data set is taken from [22] (IBIS) in the prepared form, given in [7]. The data are taken from a clinical study analyzing the response of multiple sclerosis (MS) patients to the treatment. Blood sample entrenched with mono-nuclear cells from 52 relapsing-remitting MS patients were obtained $0, 3, 6, 7, 12, 18$ and 24 months after initiation of IFNβ therapy. This resulted in 7 measurements over 2 years on average. Expression profiles were obtained using one-step kinetic reverse-transcription PCR over 70 genes selected by the specialists to be potentially related to IFNβ treatment. Overall, 8% of the measurements were missing due to patients missing the appointments. After

TABLE I
PREDICTION ACCURACIES ON THE MS TEST DATA.

Method	Number of genes	Test accuracy (%)
SGTM-TT	70	85.66 ± 8.3
SGTM-TT-R	7	93.43 ± 5.8
IBIS	3	74.20
Kalman-SVM	-	87.80
Lin-Best	7	85.00
Costa-Best	17	92.70 ± 6.1

the two year endpoint, patients were classified as either good or bad responders, depending on strict clinical criteria. Bad responders were defined as having suffered two or more relapses or having a confirmed increase of at least one point on the expanded disability status scale (EDSS). A good responder shows a suppression of relapses and does not encouter an increase of the EDSS. From the 52 patients, 33 were classified as good and 19 as bad responders. A more detailed description of the data set is available in the paper [22] and the supplemented material.

For both data sets, we use a SGTM-TT with 9 hidden states and 4 basis functions, which is a comparable model-complexity to the chosen reference methods. The analysis is done within a 4 fold cross-validation with 5 repetitions. We compare it with the general HMM classifier (HMM-Lin) and the discriminative HMM classifier (HHM-Disc-Lin) proposed in [6]. We also included the results of [22] who originally proposed the MS study, the analysis of [23], employing a Kalman Filter combined with an SVM approach and [7] proposing a semi-supervised analysis coupled with a wrapper and cut-off technique to identify discriminating features.

B. Results

1) Simulated data: We applied SGTM-TT with relevance learning for the simulated data set [6]. We observe an overall prediction accuracy of 94 ± 4. The relevance profile identified all known 10 features and effectively pruned out the remaining irrelevant data dimensions. Our results are slightly better than those reported in [6] (90%) and in [7] (92%).

2) Multiple sclerosis experiment: In Table I we summarize the prediction (test-set) results for the classification of the MS data set in comparison to the results given in [22]. We observe improved prediction accuracy employing relevance learning as compared to simple SGTM-TT. SGTM-TT improved by $\approx 6\%$ using relevance learning and the SVM classifier. Interestingly also the prediction accuracy on the full data set, including all features without relevance learning is quite good with nearly 84% and hence close to the best result proposed in [6]. The obtained mappings of the SGTM-TT are topology preserving[1] and we analyzed the mapping of the points to its prototypes and the neighborhoods. The map for the first class is depicted for two temporal sequences in Figure 4. Plots in the first row show two typical state sequences for two samples from the responder class. Also if the state sequences Z are not identical we can expect that the underlying signals X are similar due to its close neighborhood on the map. The start of a sequence is indicated by $\Box$ and the termination state by a $\circ$. This is also reflected by such similarly clustered signals at the bottom. Multiple normalized responder signals are shown using the most relevant features from Figure 5.

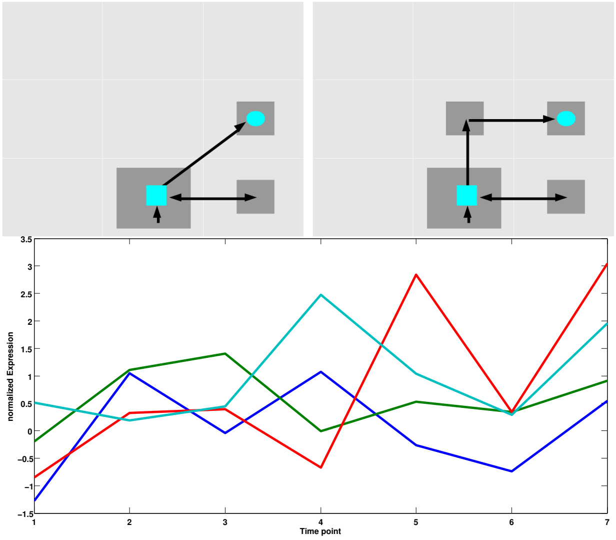


Fig. 4. Illustration of the 3×3 SGTM-TT mapping for the responder class.

As expected, results improve by integration of relevance learning compared to the full feature set. Overall the SGTM-TT with relevance learning performs very well and achieves good results of 93.43% accuracy which is comparable to the best reported model but relies on a smaller number of necessary features. The features with high relevance, are also relevant from a bio-medical point of view and show good agreement to former findings [22] by clinical experts. [2] Further the integrated relevance learning avoids multiple time consuming runs within a wrapper approach like for the techniques used in [6], [7].

The obtained relevance profile is depicted in Figure 5 and provides direct access to an interpretation of the relevant features, or marker-candidates, pruning irrelevant or noisy dimensions. We observe that the standard-deviation is relatively small, hence the relevance profiles of different runs are very stable. The most discriminative features (high-relevance), can in parts also be found in [7] but some additional features appear to be relevant, and our proposed set consists of 7 genes rather than 17 like in [7]. The values of the relevance profile are roughly Gaussian distributed with $\mu = 0.1$. We calculate a threshold ζ for the most relevant features using $\zeta = \mu + \sigma$ and obtain 7 most relevant features, summarized in Table II.

SGTM-TT also inherently models different subgroups by the probabilistic regularizing model of the GTM and GTM-TT [14], [16]. Hence the model complexity is not critical provided the map is reasonably large. This is a plus with respect to the

[1] In our observations the topographic error was reasonable small.

[2] We would like to stress that due to the small sample size and the 4 fold cross-validation a mis-classification of 1 point, accounts for an error of 8%.

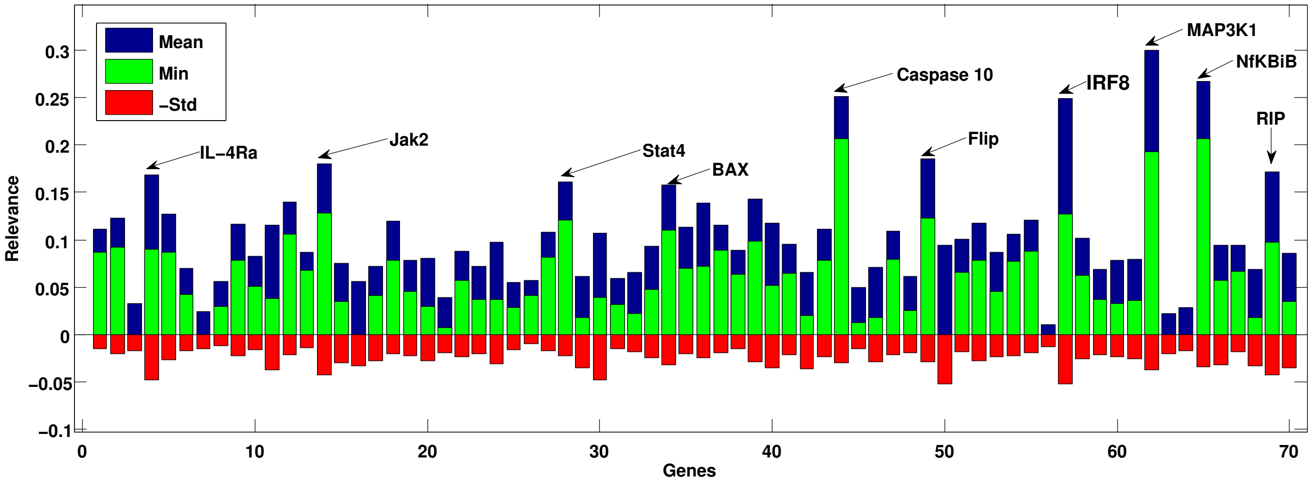


Fig. 5. Relevance profile as obtained using SGTM-TT with relevance learning. The plot shows the average relevance (blue/dark), minimal relevance (green/bright) and the standard deviation of the relevance, flipped to the negative part of the relevance axis.

TABLE II
MOST RELEVANT GENES USING SGTM-TT WITH RELEVANCE LEARNING.

Genes	Relevance	found by Lin (7)	found by Costa (17)
MAP3K1	0.3014	X	X
NFkBIB	0.2609	-	-
IRF8	0.2584	-	X
Caspase 10	0.2471	X	X
Jak2	0.1869	X	X
FLIP	0.1842	-	-
RIP	0.1647	-	-

approach presented in [7] which has the number of groups as an additional meta parameter.

VII. CONCLUSION

We have presented a novel approach for the analysis of short temporal sequences. It is based on the idea to introduce supervision and relevance learning into Generalized Topographic Mapping through time. Our results show that we are able to achieve improved or similar performance to alternative methods in the literature for a typical biomedical data set. In addition, the prototype concept of the underlying method permits a direct inspection of the model and extended visualization performance. We also obtain a direct ranking of the individual features employing the relevance profile, rather than wrapper techniques which only prune features. In future work we will explore more advanced metric adaptation schemes and alternative functional distance measures. Further we would like to apply our approach to non-clinical data.

ACKNOWLEDGMENT

The authors thank: Peter Tino, University of Birmingham for interesting discussions about probabilistic modeling and support during the early stage of this project and Falk Altheide, University of Bielefeld and Tien-ho Lin, Carnegie Mellon University, USA for support with the simulation data. We would also give extra thanks to Ivan Olier, University of Manchester, UK; Iain Strachan, AEA Technology, Harwell, UK and Markus Svensen, Microsoft Research, Cambridge, UK for providing code and support with the GTM and GTM-TT.

a) Funding:: This work was supported by the German Res. Fund. (DFG), HA2719/4-1 (Relevance Learning for Temporal Neural Maps) and by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative.

REFERENCES

[1] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.

[2] M. Strickert and B. Hammer, "Merge SOM for temporal data," *Neurocomputing*, vol. 64, pp. 39–72, 2005.

[3] T. Kohonen, *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Berlin, Heidelberg: Springer, 1995, vol. 30, (2nd Ed. 1997).

[4] I. Olier and A. Vellido, "Advances in clustering and visualization of time series using gtm through time," *Neural Networks*, vol. 21, no. 7, pp. 904–913, 2008.

[5] C. M. Bishop, "Gtm through time," in *In IEE Fifth International Conference on Artificial Neural Networks*, 1997, pp. 111–116.

[6] T. Lin, N. Kaminski, and Z. Bar-Joseph, "Alignment and classification of time series gene expression in clinical studies," in *ISMB*, 2008, pp. 147–155.

[7] I. G. Costa, A. Schönhuth, C. Hafemeister, and A. Schliep, "Constrained mixture estimation for analysis and robust classification of clinical time series," *Bioinformatics*, vol. 25, no. 12, 2009.

[8] K. M. Borgwardt, S. V. N. Vishwanathan, and H.-P. Kriegel, "Class prediction from time series gene expression profiles using dynamical systems kernels," in *Pacific Symposium on Biocomputing*, R. B. Altman, T. Murray, T. E. Klein, A. K. Dunker, and L. Hunter, Eds. World Scientific, 2006, pp. 547–558.

[9] C. Hafemeister, I. G. Costa, A. Schönhuth, and A. Schliep, "Classifying short gene expression time-courses with bayesian estimation of piecewise constant functions," *Bioinformatics*, vol. in press, 2011.

[10] J. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer, 2010.

[11] B. Hammer and T. Villmann, "Generalized relevance learning vector quantization," *Neural Networks*, vol. 15, no. 8-9, pp. 1059–1068, 2002.

[12] A. Gisbrecht and B. Hammer, "Relevance learning in generative topographic mapping," *Neurocomputing*, vol. 74, no. 9, pp. 1359–1371, 2011.

[13] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz, "Topology preservation in self-organizing feature maps: exact definition and measurement," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 256–266, 1997.

[14] C. M. Bishop, M. Svensén, and C. K. I. Williams, "Gtm: The generative topographic mapping," *Neural Computation*, vol. 10, no. 1, pp. 215–234, 1998.

[15] L. R. Welch, "Hidden Markov Models and the Baum-Welch Algorithm," *IEEE Information Theory Society Newsletter*, vol. 53, no. 4, Dec. 2003. [Online]. Available: http://www.itsoc.org/publications/nltr/it_dec_03final.pdf

[16] I. G. D. Strachan, "Latent variable methods for visualization through time," Ph.D. dissertation, University of Edinburgh, Edinburgh, UK, 2002.

[17] C. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. Springer, 2006. [Online]. Available: http://books.google.com/books?id=kTNoQgAACAAJ

[18] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl, "Regularization in matrix relevance learning," *IEEE Transactions on Neural Networks*, vol. 21, pp. 831–840, 2010.

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

[20] J. Lee and M. Verleysen, "Generalizations of the lp norm for time series and its application to self-organizing maps," in *5th Workshop on Self-Organizing Maps*, M. Cottrell, Ed., vol. 1, 2005, pp. 733–740.

[21] F.-M. Schleif, T. Riemer, U. Börner, and L. S.-H. M. Cross, "Genetic algorithm for shift-uncertainty correction in 1-D NMR based metabolite identifications and quantifications," *Bioinformatics*, vol. 27, no. 4, pp. 524–533, 2011.

[22] S. E. Baranzini, P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, X. Montalban, and J. R. Oksenberg, "Transcription-based prediction of response to ifn using supervised computational methods," *PLoS Biol*, vol. 3, no. 1, p. e2, 12 2004. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pbio.0030002

[23] R. B. Altman, T. Murray, T. E. Klein, A. K. Dunker, and L. Hunter, Eds., *Biocomputing 2006, Proceedings of the Pacific Symposium, Maui, Hawaii, USA, 3-7 January 2006*. World Scientific, 2006.