Learning relevant time points for time-series data in the life sciences

Frank-Michael Schleif, Bassam Mokbel, Andrej Gisbrecht, Leslie Theunissen, Volker Dürr, and Barbara Hammer

CITEC centre of excellence, University of Bielefeld, 33615 Bielefeld, Germany
fschleif@techfak.uni-bielefeld.de

Abstract. In the life sciences, short time series with high dimensional entries are becoming more and more popular such as spectrometric data or gene expression profiles taken over time. Data characteristics rule out classical time series analysis due to the few time points, and they prevent a simple vectorial treatment due to the high dimensionality. In this contribution, we successfully use the generative topographic mapping through time (GTM-TT) which is based on hidden Markov models enhanced with a topographic mapping to model such data. We propose an extension of GTM-TT by relevance learning which automatically adapts the model such that the most relevant input variables and time points are emphasized by means of an automatic relevance weighting scheme. We demonstrate the technique in two applications from the life sciences.

1 Introduction

Due to improved sensor technology, many data sets occurring in the biomedical domain are very high dimensional such as mass spectra or gene expression profiles. At the same time, more and more data display a temporal characteristics e.g. when investigating the development of an organism over time or the success of a therapy. In these scenarios, classical time series analysis cannot be applied due to comparably few time points (often less than 10). In addition, a direct vectorial treatment is prohibited by the high dimensionality of the data.

A few machine learning techniques exist to investigate high dimensional time series: Topographic mapping such as the self-organizing map (SOM) is extended by a recursive context which accounts for the temporal dynamics in the approach [15]. A probabilistic counterpart is offered by the Generative Topographic Mapping Through Time (GTM-TT) which combines hidden Markov models with a constraint mixture model induced by a low dimensional latent space. This approach is extended to better take the relevance of the feature components into account in [13], but relying on an unsupervised model. A supervised relevance weighting scheme which singles out relevant features in a wrapper approach based on hidden Markov models has been proposed in [12]. In [6] a similar approach introducing class-wise constraints in the hidden Markov model. The approach [5] deals with time series data and feature selection relying on support vector machines in combination with a Kalman filter. In [12], applications to life science data are presented resulting in 85% prediction accuracy on a multiple sclerosis (MS) data set, but the approach relies on strong assumptions about the

underlying HMM structure. The approach in [5] improves this result by about 3% but it results in a black box scenario without feature selection. The approach [6] is evaluated in the same scenario achieving improved performance for the MS data set. There is further ongoing work in this field, reflecting the high demand for effective methods to deal with short high dimensional time series data. The application field is not limited to the bio-medical domain [12,6,8] but covers a broader field of applications in industry and geo-science [13,15].

GTM and SOM crucially rely on the Euclidean distance which becomes more and more meaningless for high dimensional data and which suffers from an inappropriate scaling of the dimensions [11]. Because of this observation, distance based learning has been extended to automatic relevance adaptation which automatically adapts metric parameters according to given auxiliary information, see e.g. [9,7]. In this contribution, we are interested in the question how relevance learning can be transferred to the temporal domain, thereby weighting both, features and time points of the model according to their relevance as specified by given auxiliary information. The identification of relevant dimensions is very important as outlined e.g. in [13,12] to obtain a better understanding of the data, to reduce the processing complexity, and to improve the overall prediction accuracy. We propose a relevance learning scheme for GTM-TT and we demonstrate the suitability of this approach in two applications from the life sciences.

2 Generative Topographic Mapping Through Time

The Generative Topographic Mapping (GTM) as introduced in [4] models a data set X with $\mathbf{x}^i \in \mathbb{R}^D$, $i = 1, \ldots, N$ by means of a mixture of Gaussians induced by a lattice of K points $\mathbf{w}^i$ in a low dimensional latent space which can be used for visualization.

The lattice points are mapped via $\mathbf{w}^i \mapsto \mathbf{t}^i = y(\mathbf{w}^i, \mathbf{W})$ to the data space, where the function is parametrized by $\mathbf{W} \in \mathbb{R}^{m \times D}$; usually, a generalized linear regression model is chosen $y(\mathbf{w}) = \Phi(\mathbf{w}) \cdot \mathbf{W}$ with K fixed, m dimensional base functions Φ given by equally spaced Gaussians. The resulting prototypes $y(\mathbf{w}^i, \mathbf{W})$ should represent the data space as accurately as possible.

Every latent point induces a Gaussian

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

with variance β^{-1}. This gives the data distribution as a mixture of K modes $p(\mathbf{x}|\mathbf{W}, \beta) = \sum_{i=1}^{K} 1/K \cdot p(\mathbf{x}|\mathbf{w}^i, \mathbf{W}, \beta)$. Training optimizes the data log-likelihood $\ln\left(\prod_{n=1}^{N}\left(\sum_{i=1}^{K} p(\mathbf{w}^i)p(\mathbf{x}^n|\mathbf{w}^i, \mathbf{W}, \beta)\right)\right)$ by means of an expectation maximization (EM) approach with respect to the parameters $\mathbf{W}$ and β.

The GTM through time (GTM-TT) [3] extends the topographic mapping to time series data of the form $\mathbf{x} = (\mathbf{x}(1), \ldots, \mathbf{x}(T)) \in (\mathbb{R}^D)^*$ where $T \geq 1$ is the length of the time series. A data point of the training set is referred to by $\mathbf{x}^i$. Consecutive entries $\mathbf{x}^i(t)$ and $\mathbf{x}^{i+1}(t+1)$ are strongly correlated. While the space of observations over time is represented by a topographic mapping as before, the

temporal dependencies are modeled by a hidden Markov model (HMM) with hidden states characterized by the lattice points $\mathbf{w}^i$.

The HMM is parametrized by initial state probabilities $\pi = (\pi_j)_{j=1}^K$ where $\pi_j = p(\mathbf{z}(1) = \mathbf{w}^j)$ and transition probabilities $\mathbf{P} = (p_{ij})_{i,j=1}^K$ where $p_{ij} = p(\mathbf{z}(t) = \mathbf{w}^j|\mathbf{z}(t-1) = \mathbf{w}^i)$. The data probability is $p(\mathbf{x}|\Theta) = \sum_{\mathbf{z}\in\{\mathbf{w}^1,\ldots,\mathbf{w}^K\}^T} p(\mathbf{x},\mathbf{z}|\Theta)$ with parameters $\Theta = (\mathbf{W}, \beta, \pi, \mathbf{P})$, the conditional probability $p(\mathbf{x}(t)|\mathbf{z}(t)) := p(\mathbf{x}(t)|\mathbf{z}(t), \mathbf{W}, \beta)$ as before (1), and $p(\mathbf{x},\mathbf{z}|\Theta) = p(\mathbf{z}(1)) \prod_{t=2}^T p(\mathbf{z}(t)|\mathbf{z}(t-1), \mathbf{W}, \beta)$ $\prod_{t=1}^T p(\mathbf{x}(t)|\mathbf{z}(t))$ for any sequence $\mathbf{z}$ of hidden states [4].

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 [16], 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 $\mathbf{x}^n$:

$$r^{kn}(t) = p(\mathbf{z}(t) = \mathbf{w}^k|\mathbf{x}^n, \Theta) = \frac{A_{kt}B_{kt}}{p(\mathbf{x}^n|\Theta)} \tag{2}$$

with forward variables $A_{kt} = p(\mathbf{x}^n(1)\ldots\mathbf{x}^n(t), \mathbf{z}(t) = \mathbf{w}^k|\Theta)$ and backward variable $B_{kt} = p(\mathbf{x}^n(t+1)\ldots\mathbf{x}^n(t_n), \mathbf{z}(t) = \mathbf{w}^k|\Theta)$.

For an input time series $\mathbf{x}^n(1)\ldots\mathbf{x}^n(T)$, GTM-TT gives rise to a time series of responsibilities $r^{kn}(1)\ldots r^{kn}(T)$ of neuron k. Based on these responsibilities, a winner can be determined for every time step t as neuron $\mathrm{argmax}_k r^{kn}(t)$. Based on this observation, a supervised variant of GTM-TT (SGTM-TT) can be determined as follows: Assume that the time series $\mathbf{x}$ is equipped with label information l which is element of a finite set of different labels $1, \ldots L$. Then, we train a 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 prototypes $\mathbf{w}^i$. The parameters $\mathbf{W}_l$ are trained individually for every model representing label l. The same holds for the initial state probability π_l and the transition probabilities $\mathbf{P}_l$.

When processing a novel time series $\mathbf{x}$ we thus obtain L time series of responsibilities according to the labels. 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 $r_l(\mathbf{x}) := \sum_{k=1}^K \sum_{t=1}^T r_l^k(\mathbf{x}(t))/(KT)$. One can pick the label as the value l for which this quantity is largest. However, to take optimum prior class probabilities into account, we use an additional linear classifier with inputs given by the vectors $(r_l(\mathbf{x}))_{l=1}^L$ which is trained using a standard SVM.

3 Relevance learning for SGTM -TT

The principle of relevance learning has been introduced in [9] as a particularly simple and efficient method to adapt the underlying metric of prototype based classifiers according to the given situation at hand. Besides an improved data representation, it allows to interpret the relevance of the considered features to the given task. Here, we propose two different relevance weighting schemes: relevance learning of the input dimensions to change the topographic mapping according to the given class labels, and relevance learning of the time points to improve the interpretability of the results.

Relevance adaptation of the features: The squared Euclidean metric used to describe the data is substituted by the weighted form

$$d_{\boldsymbol{\lambda}}(\mathbf{x}, \mathbf{t}) = \sum_{d=1}^{D} \lambda_d^2 (x_d - t_d)^2 \,. \tag{3}$$

Relevance learning for GTM has been introduced in [7] for i.i.d. data. For SGTM-TT, a few modifications are necessary. We use the weighted metric (3) to define the Gaussians (1). 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 the relevance weighting scheme.

The question is how to set relevance parameters λ in a such way that the classification accuracy of the resulting mapping is as high as possible. We proceed as in [7] and train the relevance parameters based on priorly given class information in a separate step which is interleaved with the standard adaptation of the SGTM-TT. We rely on the cost function as introduced in generalized learning vector quantization which refers to the hypothesis margin of the classifier [14]:

$$E(\lambda) = \sum_n \mathrm{sgd}\left(\frac{d_\lambda(\mathbf{x}^n, \mathbf{t}^+) - d_\lambda(\mathbf{x}^n, \mathbf{t}^-)}{d_\lambda(\mathbf{x}^n, \mathbf{t}^+) + d_\lambda(\mathbf{x}^n, \mathbf{t}^-)}\right) \tag{4}$$

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 $\mathbf{x}^n$. sgd is the logistic function. For SGTM-TT, both, data points $\mathbf{x}^n$ and prototypes $\mathbf{t}$ are time series, the latter given by the winning prototypes of the GTM-TT model per time step. Therefore, we use a metric d_λ which constitutes a sum of the functional metric of time series components as proposed by Lee and Verleysen in [10], taking the relevance weights as parameters. Since this metric is differentiable, we can optimize this objective by means of a gradient technique.

Relevant time points: Since SGTM-TT relies on HMMs, every time point depends on its predecessor only. Thus, it is not reasonable to adapt the relevance of time points to obtain a better representation of data in the GTM-TT models. However, it is reasonable to judge the relevance of time points resulting from the GTM-TT models for the final classification, in particular if time series are of the same or a similar length. This method offers insights into which time points are particularly discriminative for the given task at hand.

We obtain a relevance profile in the following way: Denote by $r_l(\mathbf{x}(t)) := \sum_{k=1}^{K}(r_l^k(\mathbf{x}(t)))/K$ the accumulated responsibility of the GTM-TT model l for data point $\mathbf{x}^n$ at time point t. Based on this value, a classification can be based on the maximum responsibility $r_l(\mathbf{x}(t))$ in time point t. For every time point t, we simply count the number of data points which are classified correctly as belonging to class l based on the classification for time point t only, averaged over all data. A global relevance profile results thereof as sum over all labels.

4 Experiments

We consider two data sets from the biomedical domain:

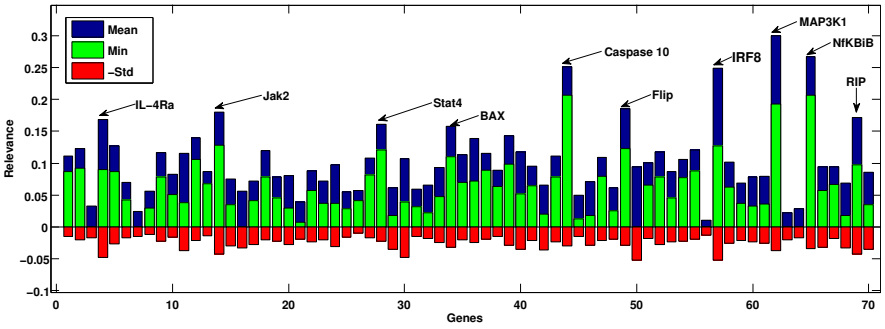

Fig. 1. Relevance profile as obtained using SGTM-TT with relevance learning. The plot shows the average relevance (blue/dark), minimal relevance (green/bright) and the standard deviation, flipped to the negative part of the relevance axis.

Multiple sclerosis data: The multiple sclerosis (MS) data set is taken from [2] (IBIS) in the prepared form, given in [6]. The data are taken from a clinical study analyzing the response of 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 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). From 52 patients, 33 were classified as good and 19 as bad responders, see [2].

We use a SGTM-TT with 9 hidden states and 4 basis functions. A 4 fold cross-validation with 5 repetitions is used. We compare the results with the general HMM classifier (HMM-Lin) and the discriminative HMM classifier (HHM-Disc-Lin) proposed in [12]. We also included the results of [2] who originally proposed the MS study, the analysis of [1], employing a Kalman Filter combined with an SVM approach and [6] proposing a semi-supervised analysis coupled with a wrapper and cut-off technique to identify discriminating features.

In Table 1 we summarize the prediction results for the MS data set in comparison to the results given in [2]. As expected, results improve by integration of relevance learning compared to the full feature set. Overall the SGTM-TT with relevance learning achieves results of 93.43% accuracy which is comparable to the best reported model but relies on a smaller number of necessary features. Further the integrated relevance learning avoids multiple time consuming runs within a wrapper approach like for the techniques used in [12,6]. The obtained relevance profile is depicted in Figure 1 and provides direct access to an interpretation of the relevant features, or marker-candidates, pruning irrelevant or noisy dimensions. The five most significant genes found by relevance learning cover three genes found by [2] and four genes found by [12,6].

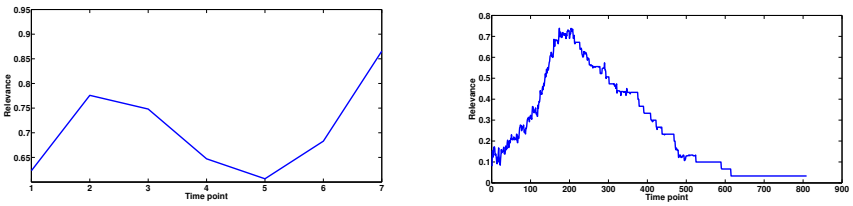

Fig. 2. Time points relevance profile for the ms data (top) and the insect data set (bottom). For the insect profile one can clearly identify a peak in the first third of the experiment, which is when insects climbed the first of two steps. The MS data indicate the most relevant time points are at $t = 2, t = 7$ with a relevance of ≈ 0.75 and ≈ 0.87 this may support a prognosis for the therapy outcome already at the second time point.

Analyzing the relevance of time points (Fig. 2) for the MS data, we observe a peak at time point two, indicating that a partial prognosis of the therapy outcome is possible already at an early stage of the therapy.

Insect locomotion data: We investigated motion-captured whole-body kinematics of insect locomotion, using data from stick insects *(Carausius morosus)*. 69 sequences where recorded in two walking conditions: a straight walk (class 1, 36% of the data), and a climbing task (class 2, 64% of the data, climbing consists of two consecutive steps of 48 mm height each). Every time step is characterized by 36 joint angles expressed in local coordinate systems of 18 leg- and 3 thorax-segments. Sequences were down-sampled from 200 Hz to 20 Hz and normalized to a standard length of 800 time steps per sequence.

An analysis of the insects data with SGTM-TT with relevance learning, 9 latent points and 4 basis functions results in a prediction accuracy of 91% percent in a 4-fold cross-validation. Relevance learning enables a 3% increase of the accuracy as compared to simple SGT-TT with 88% accuracy. The most relevant features are shown in Figure 3.

It is clearly visible that the pitch angle of the first thorax segment $T1 - y$ and the levation angles of the left legs are emphasized ($cox - y$ and $fem - y$ act synergistically). These angles display a much stronger variance when considering the climbing condition (class 2). The relevance profile of time points (see

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

Table 1. Prediction accuracies (test data) for different models using the MS data. Improved prediction accuracy employing relevance learning is observed.

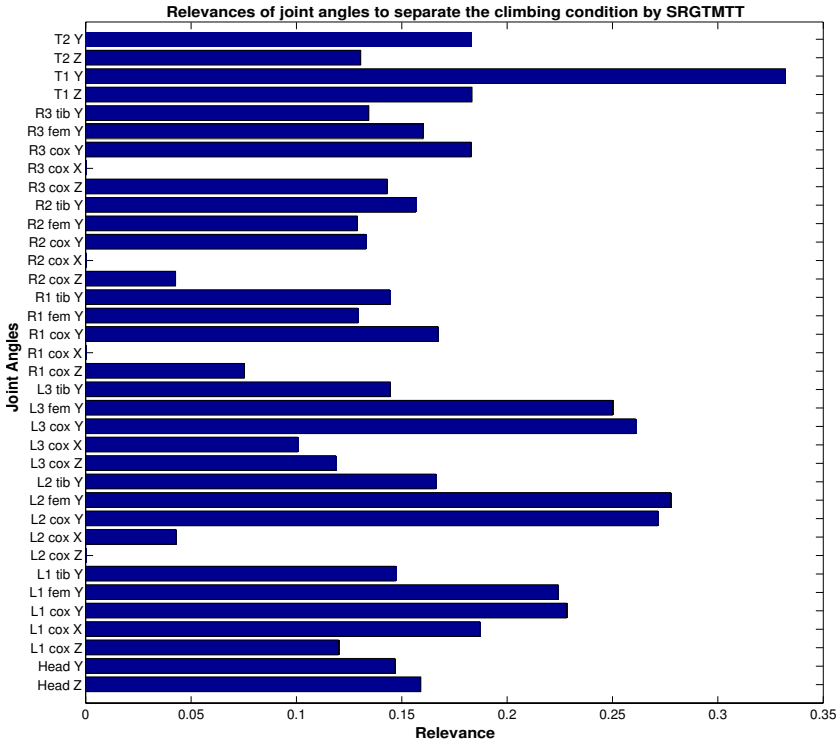

Fig. 3. Relevance profile of the joint angle features of the insect data.

Fig. 2) indicates that the most relevant time-range occurs in the first third of the dynamics, corresponding to the first ascension.

5 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 as well as a ranking of the relevance of time points. This information opens the way towards a more detailed analysis of relevant parts of the data set and the resulting model.

Acknowledgment

The authors thank: Peter Tino, University of Birmingham, for interesting discussions about probabilistic modeling 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 Manchaster, UK; Iain Strachan, AEA Technology, Harwell, UK and Markus Svensen, Microsoft Research, Cambridge, UK for providing code for GTM and GTM-TT.

Funding: This work was supported by the DFG project HA2719/4-1 to BH, by the EU project EMICAB (FP7-ICT, No. 270182) to VD, and by the Cluster of Excellence 277 CITEC funded in the framework of the German Excellence Initiative.

References

1. Altman, R.B., Murray, T., Klein, T.E., Dunker, A.K., Hunter, L. (eds.): Biocomputing 2006, Proceedings of the Pacific Symposium, Maui, Hawaii, USA, 3-7 January 2006. World Scientific (2006)
2. Baranzini, S.E., Mousavi, P., Rio, J., Caillier, S.J., Stillman, A., Villoslada, P., Wyatt, M.M., Comabella, M., Greller, L.D., Somogyi, R., Montalban, X., Oksenberg, J.R.: Transcription-based prediction of response to ifn using supervised computational methods. PLoS Biol 3(1), e2 (12 2004), http://dx.doi.org/10.1371%2Fjournal.pbio.0030002
3. Bishop, C.M.: Gtm through time. In: In IEE Fifth International Conference on Artificial Neural Networks. pp. 111–116 (1997)
4. Bishop, C.M., Svensén, M., Williams, C.K.I.: Gtm: The generative topographic mapping. Neural Computation 10(1), 215–234 (1998)
5. Borgwardt, K.M., Vishwanathan, S.V.N., Kriegel, H.P.: Class prediction from time series gene expression profiles using dynamical systems kernels. In: Altman et al. [1], pp. 547–558
6. Costa, I.G., Schönhuth, A., Hafemeister, C., Schliep, A.: Constrained mixture estimation for analysis and robust classification of clinical time series. Bioinformatics 25(12) (2009)
7. Gisbrecht, A., Hammer, B.: Relevance learning in generative topographic mapping. Neurocomputing 74(9), 1359–1371 (2011)
8. Hafemeister, C., Costa, I.G., Schönhuth, A., Schliep, A.: Classifying short gene expression time-courses with bayesian estimation of piecewise constant functions. Bioinformatics 27(7), 946–952 (2011)
9. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Networks 15(8-9), 1059–1068 (2002)
10. Lee, J., Verleysen, M.: Generalizations of the lp norm for time series and its application to self-organizing maps. In: Cottrell, M. (ed.) 5th Workshop on Self-Organizing Maps. vol. 1, pp. 733–740 (2005)
11. Lee, J., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer (2010)
12. Lin, T., Kaminski, N., Bar-Joseph, Z.: Alignment and classification of time series gene expression in clinical studies. In: ISMB. pp. 147–155 (2008)
13. Olier, I., Vellido, A.: Advances in clustering and visualization of time series using gtm through time. Neural Networks 21(7), 904–913 (2008)
14. Schneider, P., Biehl, M., Hammer, B.: Distance learning in discriminative vector quantization. Neural Computation 21, 2942–2969 (2009)
15. Strickert, M., Hammer, B.: Merge SOM for temporal data. Neurocomputing 64, 39–72 (2005)
16. Welch, L.R.: Hidden Markov Models and the Baum-Welch Algorithm. IEEE Information Theory Society Newsletter 53(4) (Dec 2003), http://www.itsoc.org/publications/nltr/it_dec_03final.pdf