ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Relevance learning for time series inspection

Andrej Gisbrecht[1], Dusan Sovilj[2],
Barbara Hammer[1], and Amaury Lendasse[2,3] *

1- University of Bielefeld - CITEC Centre of Excellence, Germany

2- Aalto University - School of Science,
Department of Information and Computer Science, Finland

3- IKERBASQUE, Basque Foundation for Science, Spain

Abstract. By means of local neighborhood regression and time windows, the generative topographic mapping (GTM) allows to predict and visually inspect time series data. GTM itself, however, is fully unsupervised. In this contribution, we propose an extension of relevance learning to time series regression with GTM. This way, the metric automatically adapts according to the relevant time lags resulting in a sparser representation, improved accuracy, and smoother visualization of the data.

1 Introduction

The self-organizing map (SOM) and variants have extensively been used for time series prediction and inspection, see e.g. [1, 9]. This is caused by the excellent generalization ability of SOM since it relies on an unsupervised prototype-based tessellation of the data space. This way, reliable models can be obtained also for high-dimensional, noisy data and few training samples. Further, SOM and alternatives such as the generative topographic mapping (GTM) offer diverse functionality besides basic regression such as the possibility to inspect typical time series motifs as expressed by the prototypes.

By far the most common way of time series prediction is based on a sliding window [5], whereby standard linear models often yield surprisingly good results. A choice of the correct time window, and, more severely, an interpretation which time lags carry meaningful information remains a challenge. Albeit an embedding usually exists due to Takens' theorem, time lags are often restricted to values derived from the autocorrelation or geometric arguments [5, 10].

Only few approaches consider a fine grained relevance detection – an information which would be beneficial for an interpretation of the dependencies present in the given time series. A couple of techniques have been proposed to extend classical prototype-based techniques such as SOM, its statistical counterpart GTM, or supervised variants to metric adaptation [6, 4, 3]. In an early attempt, this technique has been used to determine relevant dimension for classification tasks on time windows [11]. In this contribution, we transfer relevance learning to time-window based regression tasks for time series prediction based on GTM, this way arriving at an interpretable relevance profile about which time lags contribute most to a given temporal dynamics. Now, we define the basics of GTM first, introducing relevance learning into GTM, afterwards, and concluding with experiments and discussion. Thereby, time series prediction is treated as a classical regression task by a standard time window approach.

*This work has been supported by the DFG under grant number HA 2719/4-1 and by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 The generative topographic mapping

The generative topographic mapping (GTM) has been proposed as a statistical counterpart to SOM which generates data by means of a constrained mixture of Gaussians induced by a low dimensional latent space [2]. GTM models data $\mathbf{x}$ as a mixture distribution $p(\mathbf{x}|\mathbf{W}, \beta) = \sum_{k=1}^{K} p(\mathbf{w}^k) p(\mathbf{x}|\mathbf{w}^k, \mathbf{W}, \beta)$ of Gaussians

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

where $y(\mathbf{w}, \mathbf{W})$ maps prototypes $\mathbf{w}$ in low dimensional latent space to the data space by means of a generalized linear regression model induced by fixed base functions and parameterized by $\mathbf{W}$. $\mathbf{w}^k, k = 1, \ldots, K$ are the positions of lattice points in the latent space. D is the data dimensionality. Training optimizes the data log likelihood by an expectation maximization (EM) approach where the hidden variables are given by the responsibilities

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

of mode k for data point $\mathbf{x}^n$; parameters β and $\mathbf{W}$ are set in the M step, see [2].

3 Relevance learning for regression

Relevance learning has been introduced in the context of prototype based classification [4]. Essentially, it adapts the metric used to compare prototypes and data according to the relevance of the single dimensions based on auxiliary class information. It has recently been extended to unsupervised data inspection by extending GTM similarly if class labels are available [3]. Here we transfer relevance learning to GTM used for regression tasks. Assume data point $\mathbf{x}^n$ is equipped with a real valued output l^n. Posterior labeling yields the prototype label

$$c^k = \sum_{n=1}^{N} r^{kn} l^n \Big/ \sum_{n'=1}^{N} r^{kn'}, \tag{3}$$

for prototype k. N is the number of data points. This induces the smooth regression function $\mathbf{x}^n \mapsto l(\mathbf{x}^n) = \sum_{k=1}^{K} r^{kn} c^k$ and, thus, the mean squared error

$$\mathrm{MSE} = \frac{1}{N} \sum_{n=1}^{N} \left(l^n - \sum_{k=1}^{K} r^{kn} c^k\right)^2 \tag{4}$$

In relevance learning, the standard Euclidean metric in (1) is substituted by the weighted metric

$$\|\mathbf{x} - \mathbf{t}\|_\lambda^2 = \sum_{d=1}^{D} \lambda_d^2 (x_d - t_d)^2 . \tag{5}$$

The parameter λ_d indicates the relevance of dimension d. It should be adapted to emphasize the relevance of this dimension for the given regression task at

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

INIT
REPEAT
  E-STEP: DETERMINE r^{kn} BASED ON $\|\mathbf{x}-\mathbf{t}\|_\lambda^2$ AS IN (2)
  M-STEP: DETERMINE W AND β AS IN GTM ([2])
  LABEL PROTOTYPES AS IN (3)
  ADAPT λ BY STOCHASTIC GRADIENT DESCENT ON $MSE(\lambda)$ IN (4)
  NORMALIZE AND REGULARIZE λ

Table 1: Integration of relevance learning into GTM

hand. One simple way to determine λ_d is given by a minimization of the error (4). Thereby, for simplicity, we assume that the labels of prototypes c^k are fixed (which is reasonable if metric adaptation is done on a slower time scale than the adaptation of prototypes and responsibilities). In addition, we add the regularizer $\text{REG} = -\gamma/2\sum_d \exp(-\lambda_d^2)$ with parameter $\gamma \geq 0$ after gradient descent to the weights λ to enforce sparsity of the relevance profile, and we normalize the relevance profile to prevent divergence. Now the parameters λ_d can be adapted using a simple gradient technique. This metric adaptation according to a given regression task is interleaved with the standard EM training of GTM to arrive at a tessellation of the data space. A schematic representation of relevance learning can be seen in Tab. 1.

4 Experiments

We use GTM for time series prediction. Initially, a time series T is transformed using a sliding window of size D into the regression problem $\mathbf{x}^n = (T(n), T(n+1), \ldots, T(n+D-1)) \mapsto l^n = T(n+D)$. Relevance learning delivers a profile λ_d which indicates which time lag is particularly relevant within this window. We test relevance learning using three time series, see Fig. 1:

- The Laser generated data from the Santa Fe Time Series competition[1] consists of 1000 training measurements of a signal with simple short scale oscillations and a rapid decay of oscillations. The test set is of size 9093. The initial time window is of size 15.
- The El Nino data set stems from a competition organized within the Symposium ESTSP in 2007[2]. The data has 875 entries. We use a time window of 15 and evaluate the data set within a ten-fold crossvalidation.
- The Herring data set collects the seasonal spawning stock biomass of herring from the Baltic Sea main basin measured during 1974-2007. Spawning stock biomass indicates the total weight of the fish that are mature enough to reproduce. The data stems from the International Council for the Exploration of the Sea ICES[3] database associated with [7]. A total of 120 points are available. A time window 16 corresponding to a time span of 4 years is chosen. We use a leave-one-out cross-validation for evaluation.

[1] http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe.html
[2] http://research.ics.tkk.fi/eiml/datasets.shtml
[3] www.ices.dk

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

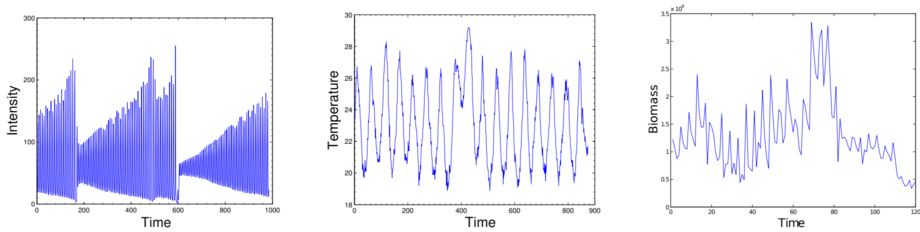

Fig. 1: Three tested time series: Sante Fe Laser data set (left), ESTSP2007 competition data (middle) and biomass of herring from the Baltic Sea (right).

GTM is initialized by the first two principal components of data. The mapping $y(\mathbf{w}, \mathbf{W})$ is based on 10×10 Gaussian base functions for Laser and El Nino and 3×3 for Herring. To speed up training, we use a 20×20 prototype grid for training for Laser and El Nino which is enlarged to 40×40 after training to achieve a better labeling resolution. Herring is trained and labeled using a 10×10 grid. The learning rate and the regularizer γ are optimized based on the MSE of the training set. This yields $\eta = 0.3$, $\gamma = 0.1$ for Laser, $\eta = 0.4$, $\gamma = 0.3$ for El Nino, and $\eta = 1.5$, $\gamma = 0$ for Herring. In comparison, we train a GTM without relevance adaptation. Additionally, an ARMA model is trained using the System Identification Toolbox. Data are z-transformed before training.

The normalized mean squared error obtained by GTM regression with and without relevance learning, as well as by ARMA, is shown in Tab. 2. Obviously, in all cases, relevance learning improves the quality of the GTM. The improved prediction accuracy is mirrored by a clearer topological ordering of the maps according to the regression task, as exemplarily depicted for the Laser data set in Fig. 3. Relevance learning reduces the topological defects with respect to the regression values. This is also confirmed by the quantization error for GTM mapping, which is 4.6744 with and 66.8650 without relevance learning. Furthermore, interesting areas of the map can be inspected by visualizing the prototypes, which depict typical behavior of the time series, see Fig. 3.

More important, relevance learning yields clear profiles which allow to infer the relevance of the given time lags, see Fig. 2. Interestingly, the variation over the cross-validation is very small concerning the judgement of several irrelevant time lags for El Nino and Herring, while it has larger fluctuation for other lags within a cross-validation. In general, these profiles confirm relevant dimensions for the two benchmark time series Laser and El Nino which have been inferred previously. For the Baltic Herring data, the relevance profile reveals interesting meaningful factors: The previous two seasons are important to be able to follow the current trend of the series. Variables 4, 8 and 12 represent the same season as the one we want to predict, as there are strong seasonal oscillations, but

	Laser	El Nino	Herring
GTM	0.0800	0.0423 (0.0017)	0.5538
Relevance GTM	0.0212	0.0368 (0.0013)	0.4252
ARMA	0.1775	0.0255	0.4202

Table 2: Normalized mean squared error for the different data sets for the test set (Laser), or a cross-validation (El Nino, Herrring)

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

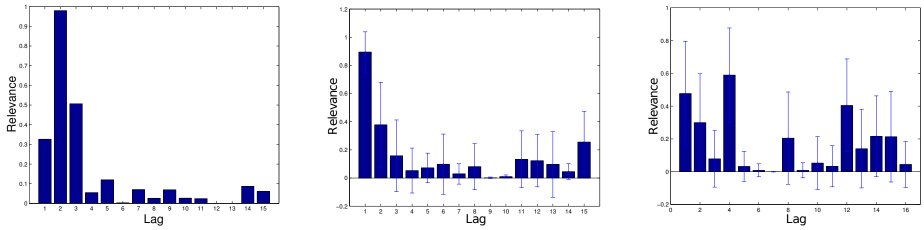

Fig. 2: Relevance profile for Laser (left), El Nino(middle), Herring (right). The standard deviation achieved in a cross-validation is also depicted for the latter.

not longer than 3 years. Interestingly, variables 13, 14 and 15 have a relevant contribution, which mirrors their relevance as the possible influx of new adults at present time, as the period of sexual maturity is around 3 years [8].

An interesting aspect of time series prediction based on an unsupervised topographic mapping is the usually resulting stable dynamics. This aspect has also been explored in the work [9]. Here, we just have a short glimpse at the long-time behavior of a GTM trained with relevance learning for the full El Nino data set. The result of a long term prediction for the next 1000 steps by means of repeated one step prediction is depicted in Fig. 4. Thereby, we use the GTM obtained while training on the full data set and the corresponding relevance profile (see Fig. 4(right)). Interestingly, only very few lags are indicated as important in this setting such that a very sparse model results. Further, the long term behavior of the time series prediction shows a very reasonable shape with periodic oscillations. Obviously, due to noise in the data, the error necessarily drops down. Nevertheless, the qualitative behavior remains stable due to its grounding in the prototype-based GTM representation.

5 Discussion

We have extended unsupervised GTM when used for regression tasks with supervised relevance learning. This way, a very intuitive model results which adapts the metric according to the given learning tasks, resulting in better interpretabil-

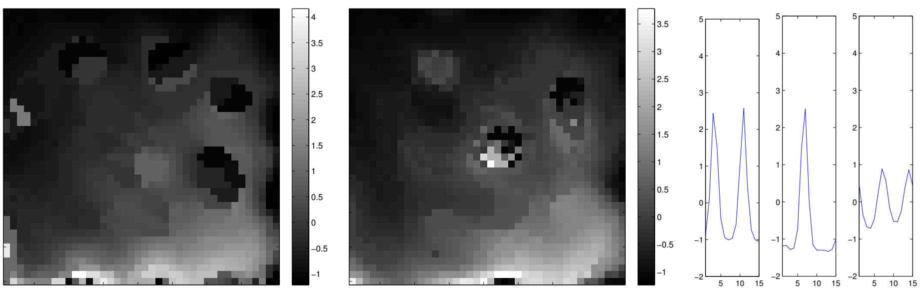

Fig. 3: GTM visualization without (left) and with relevance learning (middle) of Laser. The labeling of the prototypes is denoted as grey value. Three exemplary prototypes are depicted on the right side. The first one shows normal oscillation, the second shows the break down and the last one shows the beginning of the new oscillations.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

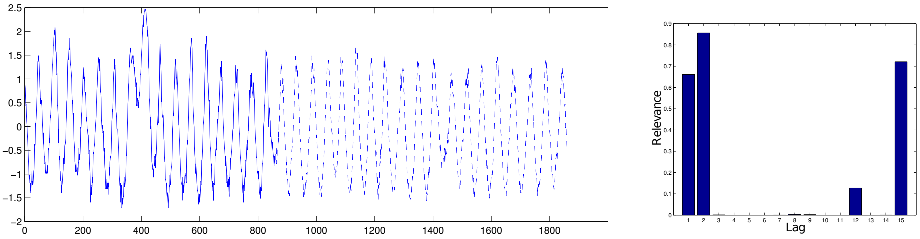

Fig. 4: Long term prediction for the El Nino data for the next 1000 time points as predicted by GTM with relevance learning (left) and relevance profile when trained on the full data set in the same setting (right).

ity and a smoother topographic mapping. We have discussed the meaning of relevance learning in the context of time window based time series prediction, and we have linked the resulting relevance profiles to relevant time lags in three exemplary benchmark data sets. This way, very sparse prototype based models result which also yield stable long term behavior. At present, the algorithmic update relies on the MSE where we assume that the prototype labeling is fixed. It remains a subject of future work to study the influence of the relevance parameters on the posterior labeling, and to complement the current update rule by rules stemming from alternative cost functions (such as e.g. multistep models).

References

[1] G. Barreto. Time Series Prediction with the Self-Organizing Map: A Review. Perspectives of Neural-Symbolic Integration, Springer, 135-158, 2007.

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

[3] A. Gisbrecht and B. Hammer: Relevance learning in generative topographic mapping. Neurocomputing 74(9): 1351-1358, 2011.

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

[5] H. Kantz and H. Schreiber. Nonlinear Time Series Analysis, 2nd edition. Springer, 2003.

[6] J. Peltonen, A. Klami, and S. Kaski. Improved learning of Riemannian metrics for exploratory analysis. Neural Networks 17(8-9): 1087-1100, 2004.

[7] Report of the ICES/HELCOM Working Group on Integrated Assessments of the Baltic Sea (WGIAB), 16-20 March 2009, Rostock, Germany. ICES CM 2009/BCC:02.

[8] M. Rajasilta. Timing of spawning in the Baltic herring (Clupea harengus membras) in the Archipelago Sea, Annales Universitatis Turkuensis. Ser. A.II. 81. Doctoral thesis, University of Turku. 1992.

[9] G. Simon, J. Lee, M. Cottrell, and M. Verleysen. Forecasting the CATS benchmark with the Double Vector Quantization method. Neurocomputing 70(13-15): 2400-2409, 2007.

[10] G. Simon, and M. Verleysen. High-dimensional delay selection for regression models with mutual information and distance-to-diagonal criteria. Neurocomputing 70(7-9): 1265-1275, 2007.

[11] M. Strickert, T. Bojer, and B. Hammer: Generalized Relevance LVQ for Time Series. ICANN 2001: 677-683.