ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 405-410

Hidden Markov gating for prediction of change points in switching dynamical systems

Stefan Liehr[*], Klaus Pawelzik

University of Bremen, Institute of Theoretical Neurophysics,
Kufsteiner Str., D–28334 Bremen, Germany
http://pooh.physik.uni-bremen.de/

Jens Kohlmorgen, Steven Lemm, Klaus–Robert Müller

GMD FIRST, Rudower Chaussee 5, D–12489 Berlin, Germany
http://www.first.gmd.de/persons/Mueller.Klaus-Robert/IDA.html

Abstract. The prediction of switching dynamical systems requires an identification of each individual dynamics and an early detection of mode changes. Here we present a unified framework of a mixtures of experts architecture and a generalized hidden Markov model (HMM) with a state space dependent transition matrix. The specialization of the experts in the dynamical regimes and the adaptation of the switching probabilities is performed simultaneously during the training procedure. We show that our method allows for a fast on–line detection of mode changes in cases where the most recent input data together with the last dynamical mode contain sufficient information to indicate a dynamical change.

1. Introduction

Non–stationarity is a severe problem in classification and prediction of dynamical systems. A framework for dealing with non–stationarity is the mixtures of experts architecture, introduced by Jacobs et al. [3]. The mixtures of experts framework aims at separating the seemingly complex global behaviour into a couple of lower dimensional sub–dynamics which can be modeled more easily. One central problem of using a set of experts is the calculation of the activities of each expert depending on the past — called the gating problem.
Many solutions have been proposed for dealing with the gating problem [1, 2, 3, 4, 7, 9, 10]. In its original formulation [3], the mixtures of experts method can be applied to systems, where different regimes do not overlap in phase space (i.e. the input space). The expert activities are provided by a feed–forward gating network given the current location in phase space [3, 10]. The use of a recurrent gating network [2] allows to distinguish also between overlapping regimes.

[*] e-mail: sliehr@physik.uni-bremen.de

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 405-410

An alternative, non–recurrent approach to distinguish between overlapping regimes is the annealed competition of experts (ACE) method [7]. It has its roots in statistical mechanics and is a purely performance–driven concept, which considers a moving average prediction error for estimating the activities instead of using a gating network.

In contrast to these well–known approaches, we use the concept of hidden Markov models (HMM) and associate each prediction expert with a hidden state of the system. Moreover, we introduce a non–linear gating network that models the conditional probabilities of transitions between the predictors depending on the actual location in phase space and the previous prediction performance. Hence, this approach unifies previous approaches by integrating (1) input information, (2) performance information, and (3) state information for modeling the gating probabilities. It is therefore also substantially more general than related HMM based methods [1, 9], which either do not make use of performance information [1] or do not use input information [9] in the gating process.

Simulation results show that mode changes can be detected much earlier, if all the three types of information are incorporated into the gating process. Likewise, the prediction performance can be improved significantly.

2. Algorithm

The concept of a generalized hidden Markov gating consists of the following three information processing components:

1. Experts: Consider a set of K models $\{f^k\}_{k=1}^K$, which can be linear or non–linear depending on the prior knowledge about the data. At time step t, $1 \leq t \leq T$, each expert provides a prediction $y_t^k = f^k(\vec{x}_t, \vec{\alpha}^k)$ which might be e.g. the estimate of a future value $y_t = x_{t+\tau}$ of a time series $\{x_t\}$ given a vector of past values $\vec{x}_t = (x_t, x_{t-\tau}, \ldots, x_{t-(d-1)\tau})$. The parameter d is called the embedding dimension and τ is called the delay parameter. Note that the extension to multivariate time series is straightforward.

The parameter vector of each model is denoted by $\vec{\alpha}^k$, the combined parameter vector of the experts is $\vec{\alpha} = (\vec{\alpha}^1, \ldots, \vec{\alpha}^K)$. Under a Gaussian assumption, the probability density r_d that a particular predictor k would have produced the data y_t, is given by

$$r_d(y_t|k) \sim \exp\left(-\beta\epsilon_t^k\right) \qquad \text{with} \qquad \epsilon_t^k = \left(y_t - y_t^k\right)^2 . \tag{1}$$

The parameter β can be interpreted as an inverse–temperature and is used for deterministic annealing during the training process.

By using Bayes' theorem this leads to the probability r_t^k that expert k has generated a given observation y_t:

$$r_t^k = r(k|y_t) = \exp\left(-\beta\epsilon_t^k\right) / \textstyle\sum_{l=1}^K \exp\left(-\beta\epsilon_t^l\right) \tag{2}$$

2. Mixing: The joint prediction y_t^* of the ensemble is given by a weighted sum of the individual outputs. Then, the probability distribution p_d for observing

y_t follows straightforward under the Gaussian assumption:

$$y_t^* = \sum_{k=1}^{K} q_t^k y_t^k \qquad p_d(y_t|\vec{x}_t, \vec{\alpha}) = \sqrt{\frac{\beta}{\pi}} \exp\left(-\beta\,(y_t - y_t^*)^2\right) \tag{3}$$

3. Gating: The mixing factors q_t^k are called the activations of each expert. They are calculated in the generalized hidden Markov gating process. In order to understand how this calculation is performed, we have to consider an HMM, which consists of (cf. [8]):
(a) a set $S = \{s^k\}$ of states, where each state is represented by a prediction expert f^k,
(b) a matrix $\hat{A}(\vec{x}_t) = \{a_t^{k|k'}\}$ of state transition probabilities, which, in our case, depend on the actual location $\vec{x}_t$ in the phase space,
(c) an observation probability distribution $p_d(y_t|s^k) = r_d(y_t|k)$,
(d) an initial state distribution $\pi = \{\pi^k\}$, which is assumed to be equally distributed.
The activation q_t^k is given by the *a priori* probability of being in state k at time t, which depends on the input–dependent transition probabilities $a_t^{k|k'}$ and the *a posteriori* state probabilities p_{t-1}^k from the previous time step:

$$q_t^k = \sum_{k'=1}^{K} a_t^{k|k'} p_{t-1}^{k'} \quad \text{with} \quad p_t^k = \frac{r_t^k q_t^k}{\sum_{l=1}^{K} r_t^l q_t^l}. \tag{4}$$

The posterior probabilities can be identified with the scaled forward probabilities $\hat{\alpha}_t^k$ in the tutorial of Rabiner [8]. They contain information about the target value y_t, whereas the prior probabilities do not.
The transition matrix is represented by an input–dependent gating network[1] h with a $(K \times K)$–output matrix and a parameter vector $\vec{\alpha}^g$: $a_t^{k|k'} = h^{k|k'}(\vec{x}_t, \vec{\alpha}^g)$. The training of experts is performed by a gradient descent on the free energy F (which is equivalent to Maximum Likelihood learning) and the gating model is optimized by using the Kullback–Leibler divergence (KL) between the posterior and prior probability distributions:

$$F = -\frac{1}{\beta T} \log \prod_t p_d(y_t|\vec{x}_t, \vec{\alpha}) \quad \text{and} \quad KL = \frac{1}{T \log K} \sum_{t=1}^{T} \sum_{k=1}^{K} p_t^k \log \frac{p_t^k}{q_t^k} \tag{5}$$

Both quantities can also be combined into one quality–function. In order to calculate the gradients of eq. (5), we use the method of Lagrange multipliers for incorporating the normalization conditions of the probabilities and the transition matrix. The training can efficiently be performed by Expectation–Maximization (EM). The E–step consists of estimating the probabilities, the M–step adapts the models by minimizing the objective functions using gradient descent.

[1] We use a radial basis function (RBF) network of the Moody–Darken type [6] for the gating network.

3. Simulations

3.1. Deterministically switching logistic map

The first example consists of a switching system of a noisy logistic map, $y_{t+1} = x_{t+1} = 4x_t(1 - x_t) + \eta_t$, and its "inverse", $y^I_{t+1} = x^I_{t+1} = 1 - 4x^I_t(1 - x^I_t) + \eta_t$, with uniform noise $\eta_t \in [-0.01, 0.01]$. The dynamics jumps from one mode to the other whenever $x_t \in [0.45, 0.55]$ holds. This is a system which exhibits non–linear behaviour, totally overlapping input spaces and a transition probability depending on the location of the dynamics. Additionally, the mean length of a mode is about only 16 time steps and therefore relatively short compared to previous applications of gated prediction systems. Modeling the jump processes is therefore very important for obtaining a high prediction quality. For the experts we use two radial basis function networks of Moody–Darken type [6] with 6 centers each, and one network of the same type but with 10 centers as the gating model.

In table 1 the performance of our method is shown compared with the ACE–algorithm [7], which uses a lowpass filter instead of modeling transition probabilities. The advantage of the HMM–based method is first a fast detection of change points, because it does not depend on a fixed smoothing algorithm. Second, the method allows a succesive iteration of the prediction system with the possibility of showing a self–driven mode change. Both properties are shown in Fig. 1.

3.2. Lorenz–System

The second example is the Lorenz system [5], which is given by a set of three coupled differential equations. With the chosen parameters the Lorenz system[2] exhibits a switching behaviour between oscillations around two fix–points. The system is globally non–linear, with the strongest non–linearity near the switch-

[2] The following parameters are used: $\sigma = 16$, $b = 4$ and $r = 45.92$.

system	algorithm	training	test 1	test 2
switching logistic map	HMM	0.00609	0.0125	0.0110
	ACE	0.00452	0.404	0.537
		$\tau_{ac} = 3$	$\tau_c = 2$	
Lorenz system	HMM	0.0270	0.0282	0.0279
	ACE	0.0287	0.0397	0.0381
		$\tau_{ac} = 9$	$\tau_c = 5$	

Table 1: Comparison of normalized mean squared errors. Note that the filter used in the ACE–algorithm is acausal on the training data (τ_{ac}) while it is causal on the test data (τ_c).

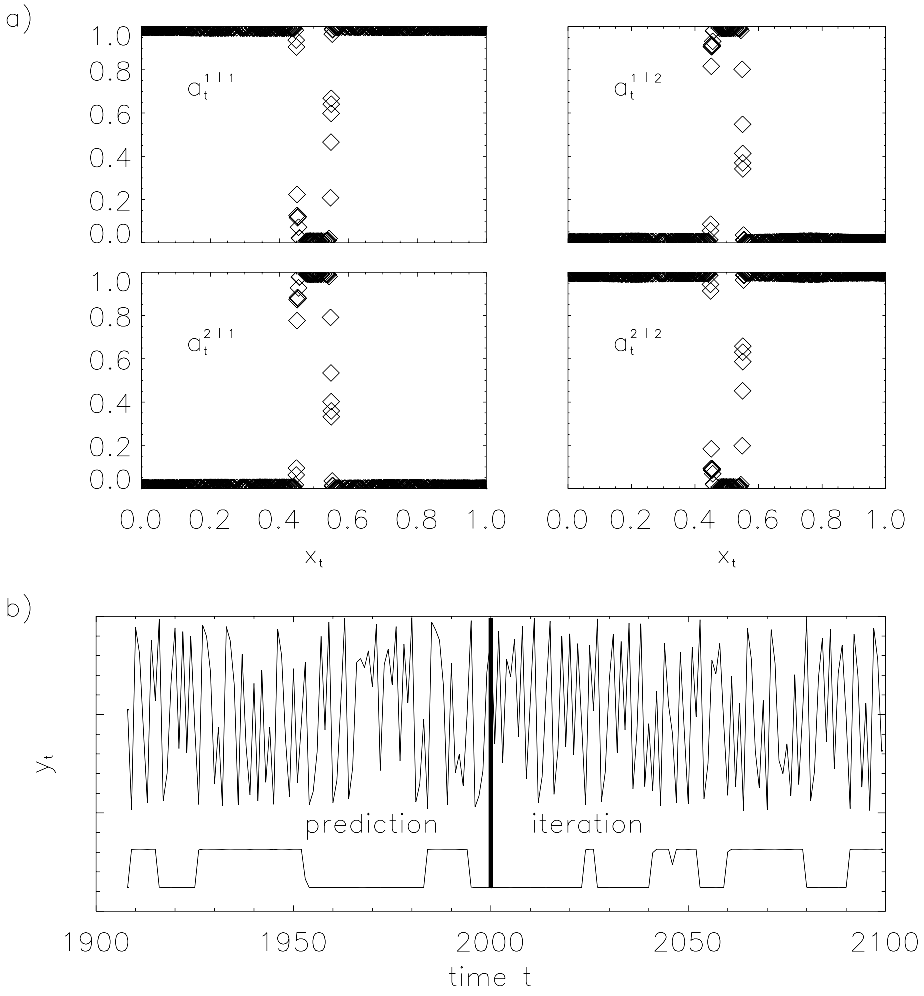

Figure 1: Hidden Markov gating of the logistic map example: a) Representation of the input dependent 2 × 2–output–matrix of the gating model. Each plot shows the one element of the transition probability matrix for changing or keeping the dynamical mode. b) One–step and iterated prediction of the time series by a gated hidden Markov mixture of non–linear predictors. Upper line: Until time step 2000 the one–step forecasts are shown, the remaining 100 time steps of the signal have been generated by iterated prediction. Lower line: The predicted allocation probabilities of one sub–dynamic. Note that the iterated prediction for $t > 2000$ includes the forecast of transitions among the modes.

ing area from one oscillatory wing to the other, while each single oscillation can be assumed to be approximately linear near the corresponding fix–point. Therefore, we choose two linear experts and a non–linear gating network for

modeling the Lorenz system. The non–linearity is thus incorporated in the gating procedure. The input and output of the experts are given by the state vector (X, Y, Z).
As shown in table 1, our algorithm yields significantly better predictions than ACE.

4. Conclusion

We presented a generalized framework for unsupervised segmentation, identification, and prediction of switching dynamics. The architecture is based on a hidden Markov model with an input–dependent state transition matrix. It consists of competing prediction experts and a gating network that, in contrast to existing methods, makes use of all available sources of information: input information from phase space, prediction error information, and HMM state information (memory). In particular, this allows for a fast detection of change points in on–line scenarios. Thereby, it can improve the prediction performance significantly. We expect that the method will be useful for the prediction of a wide range of natural signals, as e.g. climatologic or financial data.

Acknowledgement. We acknowledge support of the Deutsche Forschungsgemeinschaft (grants Pa569/2-1 and Ja379/51).

References

[1] Y. Bengio, P. Frasconi; in *NIPS'94: Advances in Neural Information Processing Systems* 7, Morgan Kaufmann, 427–434 (1995).

[2] Cacciatore, T.W., Nowlan, S.J.; in *NIPS'93: Advances in Neural Information Processing Systems* 6, Morgan Kaufmann, 719–726 (1994).

[3] R. A. Jacobs et al.; *Neural Computation* 3, 79–87 (1991).

[4] A. Kehagias, V. Petridis; *Neural Computation* 9, 1691–1710 (1997).

[5] N. Lorenz; *J. Atmos. Sci.* 20, 130 (1963).

[6] John Moody, Christian J. Darken; *Neural Computation* 1, 281–294 (1989).

[7] K. Pawelzik, J. Kohlmorgen, K.–R. Müller; *Neural Computation* 8, 340–356 (1996).

[8] L. R. Rabiner; *Readings in Speech Recognition*, A. Waibel, K. Lee, 267–296 (1988). San Mateo: Morgan Kaufmann, 1990.

[9] Shanming Shi, A. S. Weigend; *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*, 244–252 (1997).

[10] A. S. Weigend et al.; *Int. Journal of Neural Systems* 6, 373 (1995).