Improving short-term prediction with competing experts

J. Kohlmorgen†, K.-R.Müller†, K. Pawelzik‡
† GMD FIRST, Rudower Chaussee 5,
12489 Berlin, Germany
‡ Institut für Theoretische Physik, Universität Frankfurt
60054 Frankfurt/M., Germany

Abstract

We show, how competing neural networks can improve short-term prediction of time series which originate from systems with multiple modes of behaviour. With the presented method, each expert network specializes on a different dynamical mode and the time series will be segmented accordingly. In order to obtain a maximal specialization, the competition is adiabatically increased during training. Memory is included in order to resolve ambiguities of input-output relations. We illustrate the properties of the method in the case of switching chaotic dynamics. The application to Data Set D from the Santa Fe Time Series Prediction Competition demonstrates the potential relevance of this approach for time series analysis and short-term prediction.

1 Introduction

In time series prediction, neural networks can contribute substantial improvements (Weigend and Gershenfeld 1994). However, an important prerequisite for their successful application is a certain uniformity of the data: in most cases, stationarity must be assumed. If, on the contrary, a dynamical system operates in multiple modes and *switches* its dynamics, standard approaches like simple multi-layer perceptrons are likely to fail to represent the underlying input-output relations. Such time series can originate from many kinds of systems in physics, biology and engineering. Phenomena of this kind are e.g. speech (Rabiner 1988), brain data (Pawelzik 1994), and dynamical systems which switch their attractors (Kaneko 1989).

In this paper, we present a method for the prediction and segmentation of time series where no explicit information about the operating modes is given. We apply a divide-and-conquer learning strategy which forces a set of competing neural network predictors to specialize on sub-sequences of the data. Simultaneously, a segmentation according to the modes is developed.

The mixtures of experts architecture, as proposed by Jacobs et al. (1991), and also an extension, the mixtures of controllers (Cacciatore and Nowlan 1994), potentially offer a solution to this problem, since they can represent different functions by the respective experts. However, there are problems when applying these architectures to the task of identifying alternating dynamics, if the network input does not allow for a unique determination of the current mode (Pawelzik et al. 1995). We here use an ensemble of expert-networks whose competition depends only on their relative performance and *not* on the input. This way of introducing the competition is in contrast to the mixtures of experts, where an input-dependent gating network is used.

2 Annealed Competition of Experts

To illustrate the basic ideas of our approach, we discuss an example of switching chaotic dynamics. Consider a chaotic time series $\{x_t\}$ where $x_{t+1} = f_l(x_t)$, see Fig.1. Every 100 time steps the system switches to another mode l and a new map f_l, $l = 1, 2, 3, 4$, is chosen to generate the next 100 values of the time series.[1] Without knowledge about the operating mode, one cannot determine the appropriate continuation x_{t+1}, given only x_t. One way to get around this problem is the method of time-delay embedding (Packard et al. 1980). In this case, however, the inclusion of such kind

[1] The four maps are $f_1(x) = 4x(1-x), x \in [0,1]$ (logistic map), $f_2(x) = \{2x$, if $x \in [0,.5)$ and $2(1-x)$, if $x \in [.5,1]\}$ (tent map), $f_3(x) = f_1(f_1(x))$ (double logistic map), and $f_4(x) = f_2(f_2(x))$ (double tent map).

of memory, e.g. x_{t-1}, fairly complicates the prediction function. An adequate representation of the underlying relations should therefore contain a division into subtasks. Note here, that a gating network (Jacobs et al. 1991) which depends only on the input x_t must necessarily fail, whereas memory imposes a highly complicated gating task.

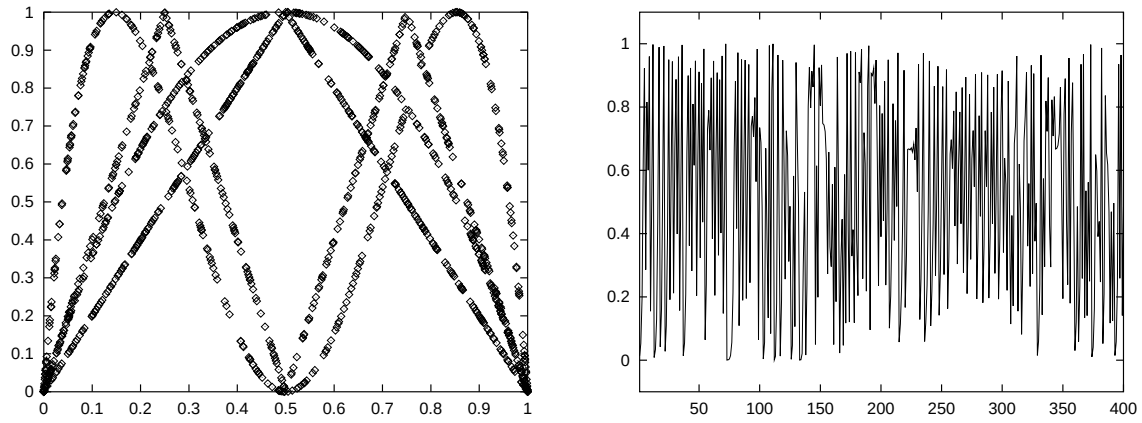


Figure 1: *(a) Training data drawn from four chaotic return maps, 300 points for each map. A new map is chosen after every 100 recursions. The first 400 values of the resulting time series are shown in (b).*

In our approach, we therefore adapt a set of predictors $\tilde{f}_i, i = 1, ..., n$, weighted only by their relative performance. The optimal choice of function approximators $\tilde{f}_i$ depends on the specific application, throughout this paper we are using radial basis function networks (RBFN's) of the Moody-Darken type (Moody and Darken 1989), because they offer a fast learning method. Under a gaussian assumption, the probability that a particular predictor i would have produced the observed data y is given by[2]

$$p(y \mid i) = Ke^{-\beta(y-\tilde{f}_i)^2} \tag{1}$$

where K is the normalization term for the gaussian distribution. Since we assume that the experts $\{i\}$ are mutually exclusive and exhaustive, we can write

$$p(y \mid \{i\}) = \sum_i p(y \mid i)p(i). \tag{2}$$

For simplicity, we further assume that the experts are equally probable, so $p(i) = 1/n$. This differs from Jacobs' mixtures of experts approach, where the $p(i)$ are produced by the gating network. In order to train the experts, we want to maximize the log-likelihood $\log L = \log(p(y \mid \{i\}))$ by a gradient ascent procedure. For the derivative with respect to the output of an expert we get

$$\frac{\partial \log L}{\partial \tilde{f}_i} \propto \left[\frac{e^{-\beta(y-\tilde{f}_i)^2}}{\sum_j e^{-\beta(y-\tilde{f}_j)^2}} \right] (y - \tilde{f}_i) \tag{3}$$

Note, that according to Bayes' rule the term in brackets is the posterior probability that expert i is the correct choice for the given data y, i.e. $p(i \mid y)$. Therefore, we can simply write

$$\frac{\partial \log L}{\partial \tilde{f}_i} \propto p(i \mid y)(y - \tilde{f}_i) \tag{4}$$

which can be interpreted as weighting the individual errors of the experts by their competence. This learning rule is also a special case of the mixtures of experts learning rule, omitting the gating network.

[2] In time series prediction we typically have $y = x_{t+1}$.

We found, however, that the learning rule in eq.(3) can be insufficient to get the correct segmentation and therewith a low prediction error. Without explicitly incorporating an assumption about the switching frequency of the dynamical modes, a variety of switching dynamical systems are conceivable as the origin of a given time series. In the current framework, the choice of models is already limited by the number of predictors, and the predictors we use only allow for relatively simple and smooth mappings. Nevertheless, it is still possible to fit the data in various ways and the training process is likely to select a wrong model and to get stuck in local minima of the error function (Fig.2(b)). Constraining the training process to find only those models with a relatively low switching rate solves the problem in cases where the dynamics does indeed switch at low rates. We do this, by imposing a low-pass filter on the errors $e_i^t = (y - \tilde{f}_i)^2$ in eq.(3):

$$E_i^t = \frac{1}{2\,\delta + 1} \sum_{t'=t-\delta}^{t+\delta} e_i^{t'} \tag{5}$$

As shown in (Pawelzik et al. 1995), the replacement of the plain errors in eq.(3) by a moving average over time, eq.(5), can formally be derived under the assumption of a low switching rate.

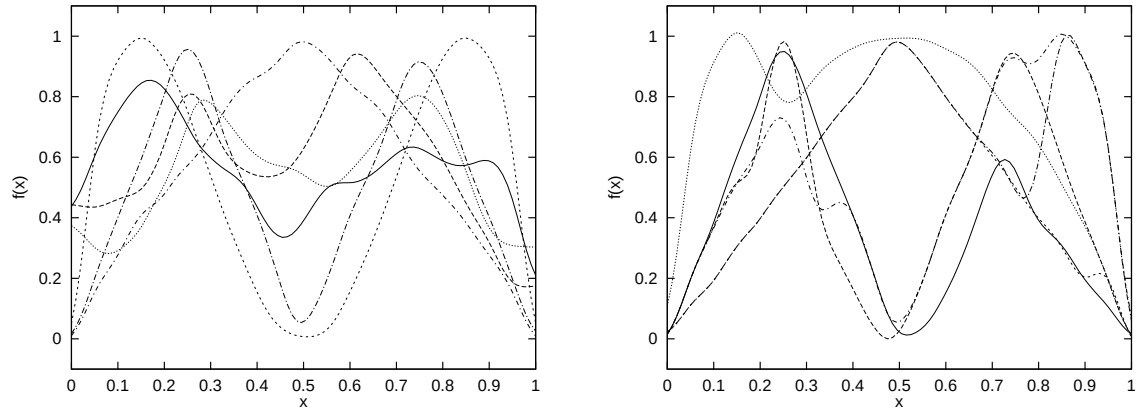
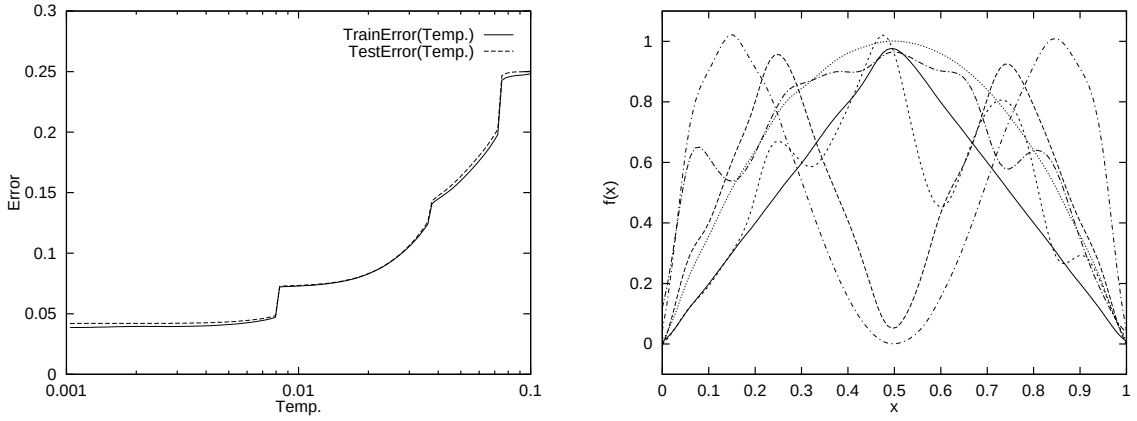


Figure 2: *(a) Result for hard competition ($\beta \to \infty$) without prior annealing: Although a proper initialization was intended, one net grabbed two "similar" return maps, f_1 and f_2. A distinction between these two maps is no longer possible and the prediction error for both maps remains high. (b) Annealing without the inclusion of memory allows the creation of maps that jump from one target map to another along the x-axis, because the information, that consecutive data points highly probable belong to the same dynamics, is not utilized.*

We also introduced an *adiabatic* annealing of the 'temperature' $T = 1/\beta$ to enable an optimal specialization of the experts, when we found that a hard competition, i.e. $\beta \to \infty$, (Kohlmorgen et al. 1994) does not always lead to a sufficient diversification of the predictors (see Fig.2(a)). We start the training process with $\beta = 0$, where the predictors equally share the same data for training. Increasing β enforces the competition, driving the predictors to a specialization on different subsets of the data. In an adiabatic annealing scheme, this diversification occurs at particular temperatures T, where the network parameters separate abruptly, resolving the underlying structure to more detail. These phase transitions are indicated by a drop of the mean prediction error (Fig.3(a)) and have been described within a statistical mechanics formalism (Rose et al. 1990). Note, that a careful decrease of T is crucial when fine differences of the underlying functions have to be resolved.

The two major enhancements, described above, have proven to be essential for a successful segmentation and prediction of time series from switching dynamics. The resulting framework also yields the correct specialization of experts in the illustrated case of switching chaotic dynamics, as shown in Fig.3(b).



(a) (b)

Figure 3: *(a) Training and test set error during the annealing process, both indicate phase transitions. (b) The maps learned by the RBFN's after training. Four nets have specialized on each of the given dynamics, while two nets dropped off and finally did not contribute to the prediction.*

3 Application to complex dynamics

We applied our method to the prediction of Data Set D from the Santa Fe Time Series Competition (Weigend and Gershenfeld 1994). This scalar data set was generated from a nine-dimensional periodically driven dissipative dynamical system with an asymmetrical four-well potential and a drift on the parameters. We used 6 RBF predictors that predict a data point using 20 preceding points, i.e. the embedding dimension was $m = 20$. The training set was restricted to the last 2000 points of Data Set D to keep the computation time tolerable. After training was finished, the prediction of the training data was shared among the predictors.

The prediction of the continuation of Data Set D was simply done by iterating the particular predictor that was responsible for the generation of the latest training data. This predicted continuation was then compared to the true one, the test set, which was originally unknown to the participants of the competition. Our method was quite useful for up to 50 time steps (see Fig.4(a)). After 50 steps, the system presumably performs a switch to another part of its potential, which per construction can not be foreseen by our approach, since the switching statistics has not been taken into account. Nevertheless, we tested the ability of this method to predict other parts of the test set by the other predictors and also found good performance up to about 50 time steps (Fig.4(b)). Again, we found that the prediction fails, when the system apparently jumps into a different state. Although the underlying system in this case was almost stationary, these results demonstrate that divide-and-conquer is a useful strategy here, because of the high dimensionality of the system and the complex form of the potential. A quantitative comparison with the winners of the Santa Fe Competition, Zhang and Hutchinson (Weigend and Gershenfeld 1994, pp. 219-241), demonstrates the power of our method. These authors applied a stationary approach that uses 100 hours of training time on a Connection Machine CM-2 with 8192 processors, and achieved a prediction error of 0.0665 (RMSE, root mean squared error), which they computed only for the first 25 step predictions, because their prediction broke down after that. Even if we compare our prediction only for this short episode, we find a RMSE of 0.0596, that is 10% better, and training took just two and a half hours on a SUN 10/20GX workstation.

4 Summary

We presented a framework for the analysis and prediction of time series. It applies to systems, where non-stationarities are caused by switching dynamics. The two major ingredients of our method are the use of a moving error-average over time and an *adiabatic* increase of the competition during training. When the method is used to predict complex dynamics, the prediction quality can be

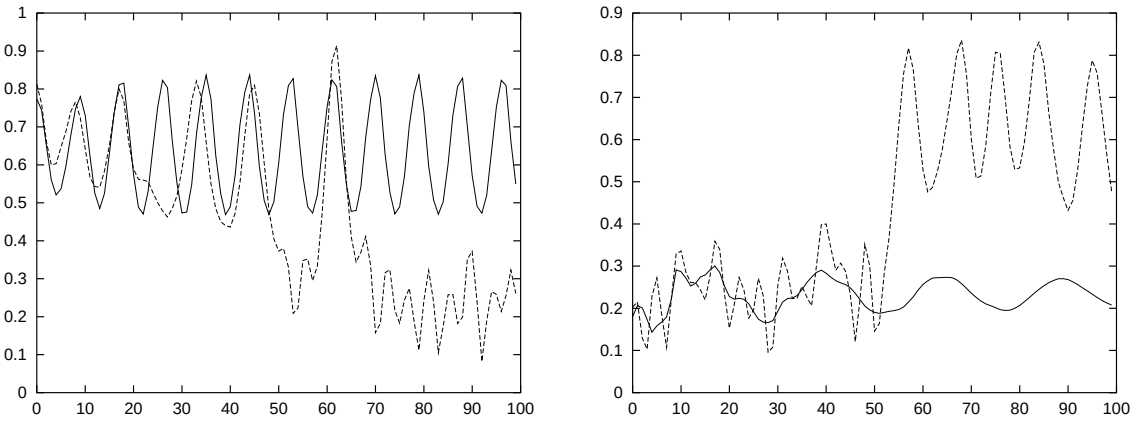


Figure 4: *Prediction (solid line) of the continuation of Data Set D (dashed line) using competing predictors. The predictors decompose the dynamics of the time series into simpler prediction tasks, so that each predictor is able to predict certain segments of the data (as shown in (a) and (b)). The accuracy for the first 25 step predictions is 10% better than the result of Zhang and Hutchinson, the winners of the Santa Fe Competition in 1991.*

improved significantly due to the divide-and-conquer strategy of the experts. In particular, we improved the results of the Santa Fe Prediction Competition (Weigend and Gershenfeld 1994) for Data Set D, which shows that this time series can efficiently be described as a switching dynamics.

Acknowledgement: K.P. acknowledges support of the DFG (grant Pa 569/1-1) and K.-R. M. acknowledges support by the European Communities S & T fellowship under contract FTJ3-004.

References

[1] Cacciatore, T.W., Nowlan, S.J. (1994). Mixtures of Controllers for Jump Linear and Non–linear Plants. NIPS'93, Morgan Kaufmann.

[2] Jacobs, R.A., Jordan, M.A., Nowlan, S.J., Hinton, G.E. (1991). Adaptive Mixtures of Local Experts, *Neural Computation* **3**, 79-87.

[3] Kaneko, K. (1989). Chaotic but Regular Posi-Nega Switch among Coded Attractors by Cluster-Size Variation, Phys. Rev. Lett. **63**, 219.

[4] Kohlmorgen, J., Müller, K.-R., Pawelzik, K. (1994). Competing Predictors Segment and Identify Switching Dynamics. ICANN'94, Springer London, pp. 1045 ff.

[5] Müller, K.-R., Kohlmorgen, J., Pawelzik, K. (1994). Segmentation and Identification of Switching Dynamics with Competing Neural Networks. ICONIP'94, Seoul.

[6] Moody, J., C. Darken (1989). Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Computation* **1**, 281–294.

[7] Pawelzik, K. (1994). Detecting coherence in neuronal data. In: Domany,E., Van Hemmen, L., Schulten, K., (Eds.), *Physics of neural networks*, Springer.

[8] Pawelzik, K., Kohlmorgen, J., Müller, K.-R. (1995). Annealed Competition of Experts for a Segmentation and Classification of Switching Dynamics, *submitted to Neural Computation.*

[9] Packard, N.H., Crutchfield J.P., Farmer, J.D., Shaw, R.S. (1980). Geometry from a Time Series. *Physical Review Letters*, 45:712-716, 1980.

[10] Rabiner, L.R. (1988). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proc. IEEE, Vol **77**, pp. 257–286.

[11] Rose, K., Gurewitz, E., Fox, G. (1990). Statistical Mechanics and Phase Transitions in Clustering. *Phys. Rev. Letters*, Vol. 65, 945–948.

[12] A.S. Weigend and N.A. Gershenfeld (Eds.) (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley.