Adaptive Memory Based Regression Methods

Hugues Bersini, Mauro Birattari, and Gianluca Bontempi

Abstract— **The task of approximating a non linear mapping using a limited number of observations, asks the data analyst to make several choices involving the set of relevant variables and observations, the learning algorithm, and the validation protocol. In the case of models which are linear in the parameters (e.g. polynomials), statistical theory and economical cross-validat ion met hods provide fast and effective ways to support these choices. However, when pure approximation performance is at stake, a unique linear structure to cover the whole range of data, is often far from optimal. Memory-based methods in contrast are well known to considerably improve the approximation performance, since all the regression analysis is done locally and repeated for each new query. In this paper, we discuss the use of these cross-validation procedures for selecting the features, the neighbors and the polynomial degree for each prediction. The possible automation of these selections on a query basis provides memory-based methods (generally not used in such a flexible way) with a larger degree of adaptivity. Experimental results in time series prediction are presented.**

Keywords— **Supervised learning, prediction.**

I. Introduction

FUNCTION approximation consists of the estimation of the regression function which assigns to each input x a number $y(x)$ equal to the conditional expectation of the scalar y (regression estimation problem [1]):

$$y(x) = \int yF(y|x)\,\mathrm{d}y = E[y|x]. \qquad (1)$$

In the classical neural approach, the problem of learning an input-output mapping is postulated as a problem of function estimation, that is of choosing from a given set of parametric functions f (x, a), a ∈ A, the one which best approximates the unknown data distribution. The problem of predicting the value of the unknown function at a new point is solved in two steps: first estimating the function, then computing the value using the estimation of the function. In this scheme one solves a relatively simple problem (estimation of the function value) by first solving a much more difficult intermediate problem (a function estimation). Function estimation is a complex task, generally treated as a minimization problem of a global cost function J (risk functional) [2] which measures the discrepancy over the whole input space between the function underlying the data set and the approximator $f(x,\alpha)$. In general, J has the following form:

$$J(\alpha) = \int L(y, f(x,\alpha))p(x,y)\,\mathrm{d}x\,\mathrm{d}y, \qquad (2)$$

where $L(y, f(x,\alpha))$ is the loss function in a point x of the input space. It is well known that when the loss-function in (2) is

$$L(y, f(X,\alpha)) = (Y - f(X,\alpha))^2, \qquad (3)$$

the regression function is the function which minimizes the functional (2) (minimum mean-square error estimator). This is essentially the perspective of the neural modeling approach. What makes this approach complex is that only a finite amount of data is available and that the risk functional has to be replaced by an approximation (empirical risk functional) constructed on the basis of the training set. Vapnik [3] showed that the minimization of the risk functional starting from the empirical risk is an ill-posed problem unless some a priori assumption about the functional dependence is made (e.g. regularization [4]).

Memory-based methods [5] aim to solve the function approximation problem taking the opposite direction: given that the problem of functional estimation is hard to be solved in a generic setting, why not to try a more restricted set of linear models and approximate the function only in the neighborhood of the point to be predicted? The main advantage of this approach comes from the fact that the simple structure of these local approximators allows an effective use of well known and powerful statistical tools. Memory-based learning turns out to be a single step approach where the whole problem is seen as a value estimation instead of a function estimation problem. This method does not build a global model of the regression function but defers any processing of the data until a new query x^* needs to be answered and then performs a linear estimation of the regression function value $E[y|x^*]$. In this case the subject of the estimation is not a function but a real value, and the data used to perform the approximation are the available samples that fall in a neighborhood of the query point. Like any other estimator inferred from a set of limited data, this approximator is affected by an error. However, unlike other approximators, this has the advantage that the assessment of its reliability is a very fast process thanks to linear cross validation or bootstrap methods. These methods make possible an iterative process which searches for the best configuration of the estimator, adapting its structure to available data.

Another important feature of this approach is the possibility to decompose a global approximation task in more simple local modeling tasks, where the hypothesis of local polynomial complexity is acceptable. This gives the data analyst the opportunity to exploit a set of theoretical results and techniques from the field of linear statistical analysis, otherwise useless in non linear domains. There are many other examples in literature where the idea of decomposition appears (e.g. regression trees [6], mixture models [7] and neuro-fuzzy inference systems [8]). How-

The authors are with IRIDIA, Université Libre de Bruxelles, 1050 Bruxelles, Belgium.

The work of G. Bontempi was supported by the European Union TMR Grant FMBICT960692.

The work of M. Birattari was supported by the FIRST program of the Région Wallonne, Belgium.

0-7803-4859-1/98 $10.0001998 IEEE

ever, unlike the memory-based approach, in these modular architectures the learning process obeys a global function estimation criterion. In this paper we will show how the adaptive memory-based methodology is not simply a different tool for better prediction, but a flexible methodology to select on-line the features, the number of relevant observations and the structure of the local model. Starting from the work of Atkeson *et* al. [9] and of Cleveland [10] on local weighted regression and from the one of Hastie and Tibshirani [11] on classification, we propose an adaptive methodology for non-linear data analysis where methods and tools from linear statistics are intensively used. We see an important field of application of this approach to problems of time series prediction where the lack of a priori knowledge about the order (i.e. the regressors) and the complexity of the model can be effectively managed with a local approach. We present some experimental results obtained in the prediction of a chaotic time series.

II. The Adaptive Memory-Based (AMB) paradigm

Modeling from data involves integrating human insight with learning techniques. In many real cases, the analyst faces a situation where a limited set of data is available and an accurate prediction is required. Often, information about the order, the structure or the set of relevant variable is missing or not reliable. The process of learning consists in a trial and error procedure during which the model is properly tuned on the available data. In a function estimation approach, the dominant criterion is the global performance of the resulting approximator over the whole input space: what is required to the model is a good performance on average. Let us consider for example an input-output mapping where the distribution of the input is not uniform. The definition of the learning problem as a risk minimization assumes that those areas of the input space where the density *p(x)* is higher deserve more attention. The risk functional, in fact, weights each prediction error $L(y, f(\text{x}, \alpha))$ according to the density value *p(x)*. As a consequence, the minimization procedure is biased towards approximators which perform better on the regions where *p(x)* is higher. On the contrary, in the memory-based approach, the estimation of the value of the unknown function is solved giving the whole attention to the region surrounding the point where the estimation is required. The classical non-adaptive memory-based procedure consists essentially in these steps:

- for each query point x^*, defining a set of neighbors, each weighted according to some relevance criterion (e.g. the dist ance);
- choosing a regression function f in a restricted family of parametric functions;
- estimating the local weighted regression;
- computing the regression value $f(x^*)$.

The data analyst who adopts a local regression approach, has to take a set of decisions related to the model (e.g. the number of neighbors, the weight function, the parametric family, the fitting criterion to estimate the parameters).

We extend the classical approach with a method that automatically selects the adequate configuration. To this aim, we simply import tools and techniques from the field of linear statistical analysis. The most important and effective of these tools is the PRESS statistic [12], which is a simple, well-founded and economical way to perform leave-one-out cross validation [13] and therefore to assess the performance in generalization of local linear models. Due to its short computation time which allows its intensive use, it is the key element of our memory-based approach to data analysis. In fact, if to each linear model, PRESS can assign a quantitative performance, alternative of models with different configurations can be tested and compared in order to select the best one. This same selection strategy is indeed exploited to select the training sub-set among the neighbors, as well as various structural aspects like the features to treat and the degree of the polynomial used as a local approximator.

A. Adaptive feature selection

A common way to deal with time series is to use a vector of time delayed observations (delay coordinate embedding [14]) to reconstruct the state space of the dynamical system underlying the time series. Following this approach, time series prediction consists in predicting a future value using time delayed observations as regressors. The search for the best set of these regressors and for their number is a major problem that have to be faced when modeling complex time series. While in the case of linear modeling a set of valuable instruments is available to deal with this problem, in the case of non-linear modeling this problem can be solved only a posteriori controlling how the global performance is sensitive to the choice of the regressors. Some of the most effective instruments that can be used in the linear case are sequential variable selection procedures (e.g. forward selection, stepwise regression, backward elimination) [12] which search for the optimal subset of regressors, or principal component analysis (PCA) techniques which compute optimal linear combination in order to avoid collinearity and reduce the dimension of the problem. With AMB it is possible to migrate these same techniques to non linear situations. The PRESS statistic turns out to be an effective way to choose which and how many regressors to use in order to improve the prediction.

B. Adaptive selection of the number of neighbors

Once the correct number and combination of regressors is chosen, the analyst must search for the best local training set consisting of the most predictive set of neighbors. Here again this selection relies on the cross-validation performance computed on sets of growing cardinality. The optimal set will be the one which presents the best intrinsic predictability in the leave-one-out sense. In our approach all neighbors are weighted by a kernel function which decreases with the distance from the query point.

C. Adaptive selection of the local polynomial degree: the possibility of fractional degree

The last parameter that is tuned automatically is the degree of the polynomial used as a local approximator. Two alternatives have been investigated. In the first one, each possible degree, from 0 (the constant model) to a given maximum, is evaluated on all the possible neighborhoods using all the possible regressors combination. A second more sophisticated alternative considers as local approximators polynomial mixings which are polynomials of fractional degree [15]. A polynomial of degree $p = m+c$, where m is an integer and $0 < c < 1$, is a linear combination of two polynomials, one of degree m and the other of degree $m + 1$, taken with weights $1 - c$ for the former and c for the latter:

$$f_p = (1-c)f_m + cf_{m+1}. \tag{4}$$

Also in this case, the value of p is selected by means of the PRESS statistics.

D. The AMB final algorithm

The general ideas of the approach can be summarized in the following way:

1. the task of learning an input output mapping is decomposed in a series of linear estimation problems;
2. each single estimation is treated as an optimization problem in the space of alternative model configurations;
3. the estimation ability of each alternative model is assessed by the cross-validation performance computed using the PRESS statistic.

This leads to the following optimization algorithm which can be described in the following form using a pseudo-programming language:

```
bestPress := Inf;
noReg is the number of regressors
for noReg := minNoReg to maxNoReg
    noNeighb is the number of neighbors
    for noNeighb := minNoNeighb to maxNoNeighb
        Deg is the polynomial degree
        for Deg := 0 to maxDeg
            if Press(noReg,noNeighb,Deg) < bestPress
                bestPress := Press(noReg,noNeighb,Deg);
                best NoRegressors : = noReg;
                bestNoNeighbors := noNeighb;
                best PolyDegree : = Deg;
            end
        end
    end
end
Prediction : = ValueEstimation($x$*,bestNoRegressors,. . .
                    bestNoNeighbors,bestPolyDegree);
```

In the future we will investigate more sophisticated search techniques than just rough exhaustive methods.

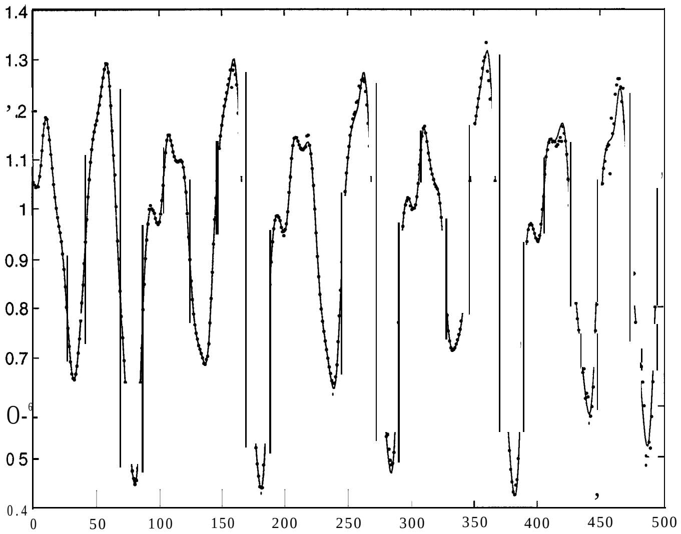

Fig. 1. Mackey Glass time series and AMB prediction (dotted line).

III. Experiments

A. Without adaptation of the number of regressors

The approach has been tested on the prediction of the chaotic Mackey-Glass time series, a well-known benchmark in time series prediction (fig. 1). We used a training set of 500 points and a test set with an equal number of samples from the benchmark available on the web’. In this first experiment we limited to the adaptive selection of the number of neighbors and the degree of the local model. We predicted the value of the series at time $t + 85$ from inputs at time $t, t-6, t-12$ and $t - 18$. We achieved a Normalized RMSE equal to 0.059. One referential result obtained with the RAN approach is NRMSE = 0.075 [16]. In fig. 2 we represent the prediction on a time window of 100 samples (diagram a), and the relative prediction square error (diagram b). Moreover, for each of the predicted time step we report in diagram (c) the optimal polynomial degree and in diagram (d) the number of neighbors taken in consideration in our iterative selection procedure. It is worth noticing how the output of the method is not simply a good prediction but a more complete information about the behavior of the dynamical system underlying the time series. In fig. 3 we report the relation existing between the square error estimated in cross-validation by the PRESS statistic (on the x-axis) and the real square error of the prediction (on the y-axis) for the prediction of one time step (the 382^{nd}). Each point in the figure represents a different model analyzed in the model search procedure. The points are roughly distributed along the line y = x. We denote with a cross (close to x = 0, y = 0) the model chosen by AMB and with a circle (close to x = 0.02, y = 0) the optimal model. Although our method selects the model represented by the point with the lowest abscissa x, while the optimal choice should be the point with the lowest or-

[1] http://legend.gwydion.cs.cmu.edu/neural-bench/benchmarks/mackey-glass.html

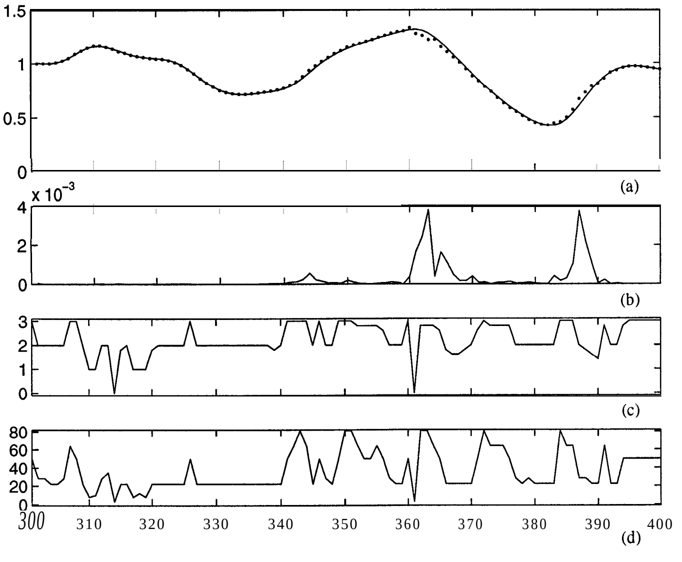

Fig. 2. Mackey Glass time series and AMB prediction (dotted line) on a **100** samples time window (a); square error (b); polynomial degree (c) and number of neighbors (d) .

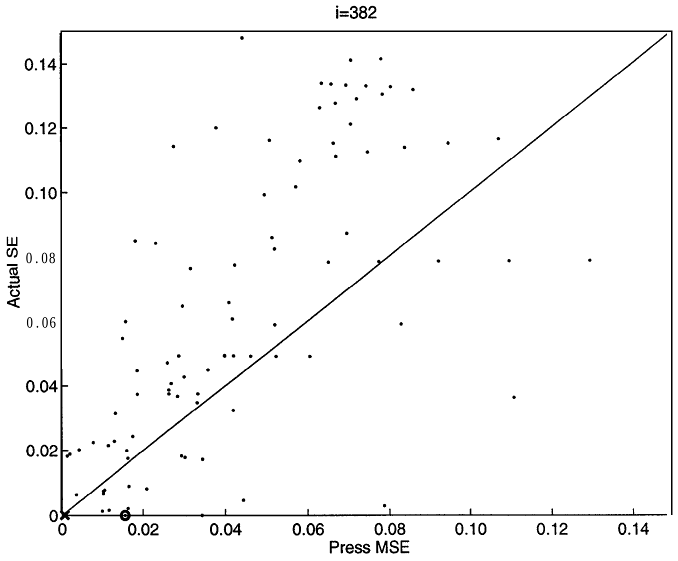

Fig. 3. PRESS estimation of the square error (z-axis) vs. real square error (y-axis).

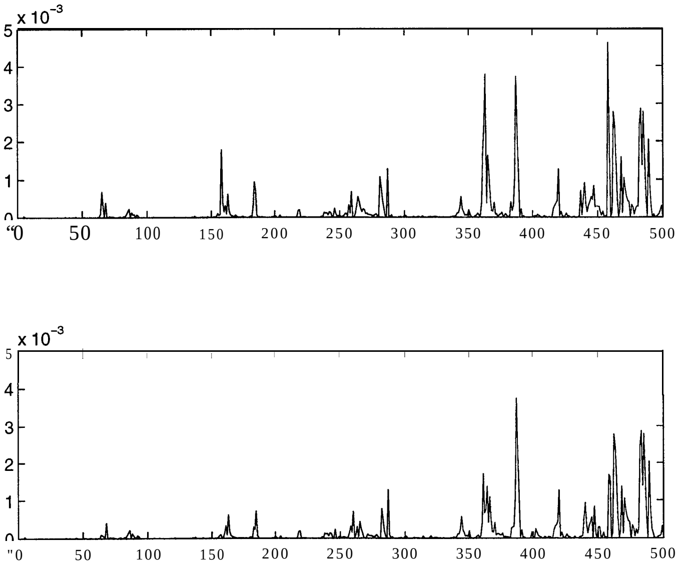

Fig. 4. Square error with a fixed number of regressors (above) and with the automatic selection procedure (below).

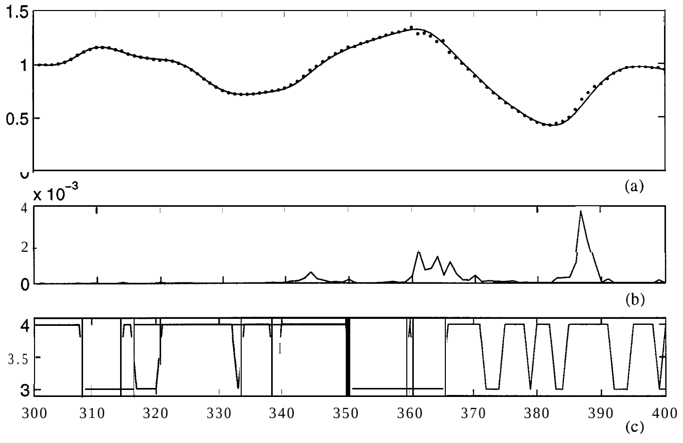

Fig. 5. Mackey Glass time series and AMB prediction with regressors selection (dotted line) on a **100** samples time window (a); square error (b) ; number of regressors (c) .

dinate, the PRESS procedure still selects a model among those with the lowest real square error. This reveals how the approach is able to have accurate and consistent predictions even in a difficult non linear configuration.

B. With adaptation of the number of regressors

Our second experiment concerns the predictions of the same chaotic time series using a AMB models which at each time step automatically selects the number of regressors that yields the most accurate prediction. In fig. 4 we report a comparison between the square error obtained with a fixed number of regressors and the error obtained with the time step selection. In this case we limited the choice between 3 and 4 regressors. We improved the previous result by achieving a Normalized RMSE equal to 0.054. In fig. 5 we report the prediction on the same time window of fig. 2, and the prediction square error (diagram b). In diagram (c) we plot the number of regressors taken into consideration for each single prediction.

IV. Conclusions

The utility of a model inferred from a limited set of available data can be evaluated according to different criteria. These criteria depend on the aim underlying the analysis of the sample. It can vary from accurate prediction to physical insight or qualitative description. The neural approach to non-linear modeling focuses mainly on the accuracy of the prediction which, due to the lack of effective tools for extracting further information from the approximator, remains essentially black-box. The adaptive memory-based learning approach, decomposing the problem into simpler sub-problems, allows to reach a comparable if not superior prediction performance and to exploit the flexibility associated with the use of linear models. It also allows the designer to obtain a more complete information about the underlying function. This information can be essential for the better exploitation of learning results which are useful for control system design or for fault diagnosis.

REFERENCES

[1] C.M. Bishop, *Neural Networks for Statistical Pattern Recognition,* Oxford University Press, 1994.

[2] V.N. Vapnik, "Principles of risk minimization for learning theory", in *Advances in Neural Information Processing Systems,* Denver, CO, 1992, vol. 4.

[3] V.N. Vapnik, *The Nature of Statistical Learning Theory,* Springer, 1995.

[4] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures", *Neural Computation,* vol. *7,* no. *2,* pp. 219-269, 1995.

[5] C.G. Atkeson, "Memory-based approaches to approximating cont inuous functions" , in *Nonlinear Modeling and Forecasting,* M. Casdagli and S. Eubank, Eds., pp. 503-521. Addison Wesley, Harlow, UK, 1992.

[6] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.

[7] M.I. Jordan and R.A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation,* vol. *6,* pp. 181–214, 1994.

[8] H. Bersini and G. Bontempi, "Now comes the time to defuzzify the neuro-fuzzy models" , *Fuzzy Sets and Sytems,* vol. 90, no. 2, pp. 161-170, 1997.

[9] C.G. Atkeson, A.W. Moore, and S. Schaal, "Locally weighted learning.", *Artificial Intelligence Review,* vol. 11, no. 1-5, pp. 11–73, 1997.

[10] W.S. Cleveland, "Robust locally weighted regression and smoothing scatterplots.", *Journal of the American Statistical Association,* vol. 74, pp. 829-836, 1979.

[11] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification and regression", in *Advances in Neural Information Processing Systems,* Cambridge, MA, 1996, vol. 8, MIT Press.

[12] R.H. Myers, *Classical and Modern Regression with Applications,* PWS-KENT, Boston, MA, 1990.

[13] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap,* Chapman and Hall, New York, NY, 1993.

[14] T. Sauer, "Time series prediction by using delay coordinate embedding", in *Time Series Prediction: forecasting the future and understanding the past,* A.S. Weigend and N.A. Gershenfeld, Eds. Addison Wesley, Harlow, UK, 1994.

[15] W.S. Cleveland, S.J. Devlin, and E. Grosse, "Regression by local fitting: met hods, properties and computation al algorithms" , *Econometrics,* vol. 37, pp. 87-114, 1988.

[16] J. Platt, "Resource-allocating network for function interpolation", *Neural Computation,* vol. 3, no. 2, 1991.