Identification of Switching Dynamical Systems Using Multiple Models

Vas. Petridis and Ath. Kehagias
Dept. of Electrical Engineering, Aristotle Univ. of Thessaloniki

Abstract

A switching dynamical system is a composite system comprising of a number of sub-systems, where, at every time step, there is a certain probability that a particular sub-system will be switched on. Identification of the composite system involves: (a) specifying the number of active sub-systems, (b) separating the observed data into groups, one group corresponding to each sub-system, (c) training a model for each subsystem and (d) combiming the subsystems to form a model of the switching system. We use the term data allocation to describe steps (a) and (b); in case accurate data allocation is available (for instance using prior information, labeled data etc.), then efficient methods are available for performing steps (c) and (d). set In this paper, however, we discuss the case where data allocation is not available and steps (a) and (b) must be performed concurrently with (c) and (d). This is, essentially, a problem of unsupervised learning. We present here conditions sufficient to ensure the convergence of a quite general class of data allocation schemes and relate these conditions to PAC learnability. The theoretical conclusions are supported by numerical experiments on a problem of on-line switching system identification.

1 Introduction

Consider a dynamical system of the form

$$y(t) = f_{z(t)}(y(t-1),\ldots,y(t-M),u(t),\ldots,u(t-N)); \quad (1)$$

where $y(t)$ is the observation, $u(t)$ is the input and $z(t)$ is the switching variable of the dynamical system. "Switching" means that, at every time step t, $z(t)$ takes a value in the finite set $\{1,2,\ldots,K\}$ and for each such value we a sub-system of the form

$$y(t) = f_k(y(t-1),\ldots,y(t-M),u(t),\ldots,u(t-N)); \quad (2)$$

(where k takes the values $1,2,\ldots,K$) is activated. Each of the K systems of the form of eq.(2) is a NARMAX system [1]; state space systems can be transformed in the NARMAX [1] form of eq.(2). The switching system of eq.(1) is more general than either NARMAX or state space systems. In what follows we will sometimes use the term "source" as equivalent to "sub-system" and we will say that "$y(t)$ is generated by the k-th source", meaning that at time t the switching variable $z(t)$ takes the value k.

In order to model the system of eq.(1), it is natural to utilize a collection of models of the form

$$\hat{y}_k(t) = \hat{f}_k(y(t-1),\ldots,y(t-\hat{M}),u(t),\ldots,u(t-\hat{N})); \quad (3)$$

where $k = 1,2,\ldots,K.$, and the functions $\hat{f}_k(\cdot)$ are black-box models (e.g. neural networks, fuzzy systems and so on). Assuming that for each k a model $\hat{f}_k(\cdot)$ which approximates $f_k(\cdot)$ well (in an appropriate sense) is available, efficient methods [2, 3, 4, 5] exist for combining the K models of eq.(3), so as to form a composite model which accurately models the input-output behavior of the true system of eq.(1). Furthermore, it is well known that many classes of black-box models are universal approximators [7, 8], hence it is not difficult to obtain good models of the form of eq.(3), provided that training data are available which correspond to each of the K subsystems.

However, in an online, adaptive system identification context, an incoming stream $y(1), y(2), \ldots, y(t), \ldots$ of unlabeled data will be available, and it is not immediately obvious how to allocate these between the available models. In other words, the source which generated $y(t)$ is not known a priori. For any given data allocation scheme, it is likely that, in the initial stages of training, data generated by the k-th source may be allocated to several models; conversely, each model may receive data generated by several distinct sources. What is required for succesful identification is specialization: eventually every model should consistently accept data corresponding to a particular source and reject all other data; if this occurs then we say that the data allocation scheme converges. This implies that the number of active sources is discovered, as well.

In this paper sufficient conditions are given for the convergence of a general class of data allocation schemes. It is proved that, under quite general and reasonable conditions, a data allocation scheme will converge in the sense previously discussed: exactly one model will specialize to each active source.

2 Example: Data Allocation by a Predictive Performance Criterion

Consider the following recursive online algorithm as an example of a scheme for data allocation and training of K models. The algorithm is serial; that is, an incoming datum is tested against the first model, if it is not considered appropriate for the first model is test against the second model and so on.

Serial Data Allocation Scheme

```
1. Initialization: Set $K$ equal to 1. Set a threshold $\varepsilon$. Initialize randomly a model $\hat{f}_1^{(0)}(\cdot)$.
2. For $t = 1, 2, \ldots$ do the following
   (a) For $k = 1, 2, \ldots, K$ compute estimates $\hat{y}_k(t)$ of $y(t)$, using the models $\hat{y}_k(t) = \hat{f}_k^{(t-1)}(y(t-1), y(t-2), \ldots, y(t-\hat{M}), u(t), u(t-1), \ldots, u(t-\hat{N}))$.
   (b) Observe $y(t)$.
   (c) For $k = 1, 2, \ldots, K$ : if $|y(t) - \hat{y}_k(t)| \le \varepsilon$, set $\hat{z}(t) = k$ and break out of the loop. If $|y(t) - \hat{y}_k(t)| > \varepsilon$ for $k = 1, 2, \ldots, K$, then set $\hat{z}(t) = K + 1$ and increase $K$ to $K + 1$.
   (d) Allocate $y(t)$ to model nr.$\hat{z}(t)$.
   (e) For $k = 1, 2, \ldots, K$: retrain the $k$-th model, using all data so far allocated to it, to obtain a new model $\hat{f}_k^{(t)}(\cdot)$..
3. Next $t$.
```

Prima facia, this scheme has a good chance of successful specialization. If a model has, at some stage of the algorithm, collected enough data generated by the k-th source, assuming efficient training, it will be a good model of the k-th source behavior. In a later occurence of $z(t) = k$, the same model will be likely to produce a further good estimate $\hat{y}_k(t)$ and hence accept $y(t)$ in its training data pool. On the other hand, if $z(t) = m$ (with $m \neq k$), then (assuming the K sources have sufficiently distinct input/output behavior) the same model will be likely to produce a poor estimate $\hat{y}_k(t)$ and hence pass $y(t)$ for examination by the remaining models. As t goes to infinity, the model collects predominantly data generated by the k-th source and, if the training scheme is efficient, $\hat{f}_k^{(t)}(\cdot)$ becomes a progressively better model of $f_k(\cdot)$.

We have considered the particular data allocation scheme, which depends on a predictive accuracy criterion, merely for purposes of illustration. In fact, using the same argument as above, we can expect succesful specialization for any data allocation scheme that has the following property: whenever a model accepts a datum generated by the k-th source, then the probability of the same model accepting further k-th source generated data increases, while the probability of the same model accepting m-th source generated data (with $m \neq k$) decreases. In the next section we present the same argument in a more mathematical form, and provide sufficient conditions to ensure succesful specialization of the data allocation scheme.

3 Convergence Analysis

We present the convergence theorems we have obtained, but omit the proofs, for reasons of brevity. The interested reader may find the proofs in [6].

3.1. Two Sources. We start the convergence analysis by considering the simplest possible switching dynamical system; this is described by eq.(1) with $K = 2$. I.e. $z(t)$ takes values in the set $\{1, 2\}$. Suppose that $z(t)$ is an i.i.d. sequence, with $\Pr(z(t) = i) = \pi_i$, $i = 1, 2$. Obviously $\pi_1 + \pi_2 = 1$; it is also assumed that

B1 for $i = 1, 2$ we have $0 < \pi_i < 1$.

The estimate $\hat{z}(t)$ also takes values in $\{1, 2\}$; $\hat{z}(t) = i$ means that $y(t)$ has been allocated to the i-th model. Consider also the variables $M_{ij}(t)$ (where $t = 1, 2, \ldots$ and $i, j = 1, 2$) defined by

$$M_{ij}(t) = \begin{cases} 1 & \text{if} \quad z(t) = i, \hat{z}(t) = j \\ 0 & \text{else;} \end{cases}$$

and the variables $N_{ij}(t)$ (where $t = 1, 2, \ldots$ and $i, j = 1, 2$) defined by

$$N_{11}(t) = \sum_{s=1}^{t} M_{11}(s); \qquad N_{12}(t) = \sum_{s=1}^{t} M_{12}(s);$$
$$N_{21}(t) = \sum_{s=1}^{t} M_{21}(s); \qquad N_{22}(t) = \sum_{s=1}^{t} M_{22}(s).$$

In other words $N_{ij}(t)$ indicates the total number of type i samples assigned to model j, up to time t. Now consider the variable $X(t)$ defined by

$$X(t) = N_{11}(t) - N_{21}(t).$$

$X(t)$ is the specialization variable. If $X(t)$ is large and positive, then model nr.1 has received a large surplus of data generated by source nr.1. Similarly, in case $X(t)$ is large and negative, model nr.2 has received a large surplus of data generated by source nr.2. In short, a large value of $|X(t)|$ indicates that model nr.1 has specialized on one of the two sources. Now, define data allocation probabilities

$$
\begin{aligned}
f(n) &= \Pr(\widehat{\ }(t) = 1 \mid z(t) = 1, X(t-1) = n);\\
g(n) &= \Pr(\widehat{\ }(t) = 1 \mid z(t) = 2, X(t-1) = n).
\end{aligned}
$$

By the preceding argument, it is reasonable to make the following assumptions, regarding these probabilities

A1 For $n = 1, 2, \dots$, $f(n) > 0$ and $\lim_{n \to -\infty} f(n) = 0$; $\lim_{n \to +\infty} f(n) = 1$;

A2 For a $n = 1, 2, \dots$, $g(n) > 0$ and $\lim_{n \to -\infty} g(n) = 1$; $\lim_{n \to +\infty} g(n) = 0$.

A1 says that, when local model nr.1 has already specialized on source nr.1, then it will very likely accept an additional source 1 generated datum and will very unlikely reject an additional source 2 generated datum. Similar remarks can be made regarding A2.

By the description of the data assignment procedure it follows that $X(t)$ is a Markovian process. The transition probabilities (for $m, n \in Z$) are obtained from the data allocation method and are

$$
\begin{aligned}
p_{n,m} &= 0 \text{ if } |n-m| > 1;\\
p_{n,n-1} &= \pi_2 \cdot g(n);\\
p_{n,n+1} &= \pi_1 \cdot f(n);\\
p_{n,n} &= \pi_1 \cdot (1 - f(n)) + \pi_2 \cdot (1 - g(n)).
\end{aligned}
$$

The following two theorems describe the convergence behavior of a data allocation scheme that satisfies conditions A1, A2, B1. The first theorem describes the behavior of $X(t)$.

Theorem 1 If conditions B1, A1, A2 hold, then

(i) $\forall m = 0, \pm 1, \pm 2, \dots \quad \Pr(X(t) = m \text{ i.o.}) = 0$;

(ii) $\Pr\left(\lim_{t \to \infty} |X(t)| = +\infty\right) = 1$;

(iii) $\Pr\left(\lim_{t \to \infty} X(t) = +\infty \text{ or } \lim_{t \to \infty} X(t) = -\infty\right) = 1$.

In this theorem, the most important conclusion is (iii): if conditions B1, A1 and A2 hold, then at least one of the two models will (in the long run) accumulate either a lot more source 1 generated data than source 2 generated data ($X(t) \to +\infty$) or the other way round ($X(t) \to -\infty$). If $X(t) \to +\infty$, then at least one of the two models will specialize (either model nr.1 in source 1 or model nr.2 in source 2). Conversely, if $X(t) \to -\infty$, then then at least one of the two models will specialize. The total probability that one of these two events will take place is one, i.e. at least one model will certainly specialize in one of the two sources.

In fact, however, Theorem 1 is used as stepping stone to prove that both local models will specialize, each in a different source, and in a stronger sense. This is stated in the next theorem.

Theorem 2 If conditions B1, A1, A2 hold, then

1. If $\Pr(\lim_{t \to \infty} X_t = +\infty) > 0$ then

$$
\Pr\left(\lim_{t \to \infty} \frac{N_t^{21}}{N_t^{11}} = 0 \,\middle|\, \lim_{t \to \infty} X_t = +\infty\right) = 1;
$$

$$
\Pr\left(\lim_{t \to \infty} \frac{N_t^{12}}{N_t^{22}} = 0 \,\middle|\, \lim_{t \to \infty} X_t = +\infty\right) = 1.
$$

2. If $\Pr(\lim_{t \to \infty} X_t = -\infty) > 0$ then

$$
\Pr\left(\lim_{t \to \infty} \frac{N_t^{11}}{N_t^{21}} = 0 \,\middle|\, \lim_{t \to \infty} X_t = -\infty\right) = 1;
$$

$$
\Pr\left(\lim_{t \to \infty} \frac{N_t^{22}}{N_t^{12}} = 0 \,\middle|\, \lim_{t \to \infty} X_t = -\infty\right) = 1.
$$

Theorem 2 states that, with probability one, both predictors will specialize, one in each source and in a "strong" sense . For instance, if $X(t) \to +\infty$, then the proportion $\frac{N_{21}(t)}{N_{11}(t)}$ (nr. of source 2 samples divided by nr. of source 1 samples assigned to predictor 1) goes to zero ; this means that "most" of the samples on which predictor 1 was trained come from source 1 and, also, that "most" of the time a sample of source 1 is assigned (classified) to the predictor which is specialized in this source. Hence we can identify source 1 with predictor 1. Furthermore the proportion $\frac{N_{12}(t)}{N_{22}(t)}$ (nr. of. source 1 samples divided by nr. of source 2 samples assigned to predictor 2) also goes to zero ; this means that "most" of the samples on which predictor two was trained come from source 2 and, also, that "most" of the time a sample of source 2 is assigned (classified) to the predictor which is specialized in this source. Hence we can identify source 2 with predictor two. A completely symmetric situation holds when $X(t) \to -\infty$. By Theorem 1, $X(t)$ goes either to $+\infty$ or to $-\infty$, so specialization of both predictors (one in each source) is guaranteed.

3.2. Many Sources. Now consider the switching dynamical system of eq.(1) with $K > 2$. I.e. we have K sub-systems ("sources") and $z(t)$ takes values in the set f1; 2; :::; Kg. Consider now a new variable æ(t) taking values in f1; 2g according to the following rule

$$\text{æ}(t) = \text{½} \begin{matrix} 1 & \text{i¤ } z(t) = 1; \\ 2 & \text{i¤ } z(t) > 1: \end{matrix}$$

This new variable corresponds to two sources: the first source is the actual source nr.1 and the second is a composite source comprising of all the remaining sources. If conditions B1, A1, A2 hold with respect to the new variable æ(t) then, by Theorems 1 and 2, in the long run one model will specialize in the simple source and the other model will specialize in the composite source. Now consider a new data set, comprising of the data allocated to the composite source. Also consider a new source set comprising of the simple source nr.2 and a new composite source consisiting of simple sources f3; 4; :::; Kg. The data allocation algorithm can be applied on the new data set; if conditions B1, A1, A2 hold with respect to the new source set (for this to be true it may be required to lower the threshold ") then one new model will specialize on simple source nr.2 and another model will specialize on composite source f3; 4; :::; Kg: Continuing in this manner, K models can be obtained, each approximating one simple source.

4 A Connection to PAC Learnability

In case we limit ourselves to data allocation schemes which utilize a predictive accuracy criterion, we can restate the convergence conditions A1 and A2 in a form which relates to PAC learnability[9]. Denote the prediction error of model nr.1 by $e_1(t)$, i.e.

$$e_1(t) = jy(t) \text{ ¡ } \text{þ}_1(t)j:$$

Now define probabilities

$$\begin{aligned} F(n; \text{"}) &= Pr(e_1(t) < \text{"} jz(t) = 1; X(t \text{ ¡ } 1) = n); \\ G(n; \text{"}) &= Pr(e_1(t) > \text{"} jz(t) = 2; X(t \text{ ¡ } 1) = n): \end{aligned}$$

Consider the task (assigned to model nr.1) of recognizing source 1 data. This task is effected succesfully (for $z(t) = 1$) exactly when $e_1(t) <$ ". The probability that this task is completed with success at time t (given that the threshold used is " and that model 1 has already accepted n more source 1 data than source 2 data) is exactly F(n; "). Now, suppose that this task is PAC learnable. This is equivalent to:

$$\begin{aligned} \text{for all } \text{±} &> 0; \text{"} > 0 \text{ exists some } n_0 \text{ such that} \\ \text{for all } n &> n_0 \text{ we have } F(n; \text{"}) > 1 \text{ ¡ ±}; \end{aligned} \quad (4)$$

But F(n; ") is exactly equal to f(n) (for the particular " used) and eq.(4) is exactly equivalent to condition A1. In short, if the task of source 1 data recognition is PAC learnable, then condition A1 is satisfied.

Similarly, consider the task (assigned to model nr.1) of rejecting source 2 data. This task is effected succesfully (for $z(t) = 2$) exactly when $e_1(t) >$ ". The probability that this task is completed with success at time t (given that the threshold used is " and that model 1 has already accepted n more source 1 data than source 2 data) is exactly G(n; "). Now, suppose that this task is PAC learnable. This is equivalent to:

$$\begin{aligned} \text{for all } \text{±} &> 0; \text{"} > 0 \text{ exists some } n_0 \text{ such that} \\ \text{for all } n &> n_0 \text{ we have } G(n; \text{"}) > 1 \text{ ¡ ±}; \end{aligned} \quad (5)$$

which can be rewritten as

$$\begin{aligned} \text{for all } \text{±} &> 0; \text{"} > 0 \text{ exists some } n_0 \text{ such that} \\ \text{for all } n &> n_0 \text{ we have } 1 \text{ ¡ } G(n; \text{"}) < \text{±}; \end{aligned} \quad (6)$$

But 1 ¡ G(n; ")= g(n) (for the particular threshold " used) and eq.(6) is exactly equivalent to condition A2. In short, if the task of source 2 data rejection is PAC learnable, then condition A2 is satisfied. In conclusion, if the acceptance and rejection tasks are PAC learnable, then conditions A1, A2 will hold and the data allocation process will be succesful.

5 Experiments

In this section a simple data allocation algorithm is applied to several problems of switching dynamical systems identification. Two sets of experiments are presented. Data allocation is performed by the algorithm presented in Section 3.

A. In experiment group A, two chaotic dynamical subsystems are used, as follows:

1. for $z(t) = 1$, a logistic time series of the form $y(t) = f_1(y(t \text{ ¡ } 1))$, where $f_1(x) = 4x(1 \text{ ¡ } x)$;
2. for $z(t) = 2$, a tent-map time series of the form $y(t) = f_2(y(t \text{ ¡ } 1))$, where $f_2(x) = 2x$ if x 2 [0; 0:5) and $f_2(x) = 2(1 \text{ ¡ } x)$ if x 2 [0:5; 1];

The two sub-systems are activated alternately, each for 200 time steps, resulting in a period (for the z(t) process) of 400 time steps. Ten such periods are used, resulting in a 4000-steps time series. The task is to discover the two sub-systems and the activation schedule, as well as to develop one neural network model for each system. The data allocation algorithm is used with K = 2, i.e. with two neural networks; 1-4-1 neural networks are used (i.e. one input, four hidden neurons and one output). The switching dynamical system is observed at various levels of noise, i.e. at every step y(t) is mixed with additive white noise, distributed uniformly in the interval [¡A=2; A=2]. To evaluate the quality of system identification the figures c and d are used. For the computation of c, classifications from t = 3001 until t = 4000 are taken into account, i.e. when both sub-systems have been learned by the models; we have $c = T_0/1000$, where T_0 is the number of correctly classified time steps after t = 3001. The prediction error (an index of how good are the developed models) is computed according to the formula $d = \sqrt{\frac{\sum_{t=3001}^{4000} \left| y(t) - \hat{y}_{\hat{z}(t)}(t) \right|^2}{\sum_{t=3001}^{4000} |y(t)|^2}}$. The experiment is repeated six times for every noise level and the c and d figures found for every experiment are averaged and presented in Figure 1.

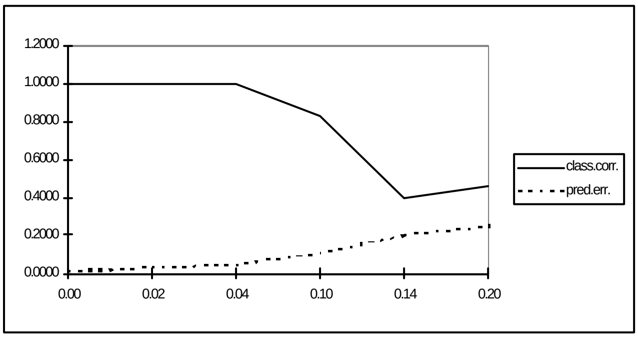

Figure 1: Result of Experiment Group A. Classification figure of merit c and prediction error d. Solid line corrsponds to c; dotted line corresponds to d. The horizontal axis indicates noise level A.

It is seen that for noise levels up to A = 0:10, data allocation is successful; after that point it gradually deteriorates. Note that performance deterioration takes place at high noise levels, where two effects take place: (a) input / output behavior is not easily distiguinshable for the two sub-systems and (b) prediction is quite poor. Hence, two factors enter which may lead to a violation of conditions A1 and A2. In this sense the importance of these conditions for convergence is corroborated. A representative profile of z(t) (for time steps t = 1; 2; :::; 1000) is presented in Figure 2 and one for prediction error $e(t) = y(t) - \hat{y}_{\hat{z}(t)}(t)$ (for time steps t = 1001; 1002; :::; 1200) in Figure 3; both of these correspond to the noise free case. In Figure 2 note that for t = 1; 2; :::; 200 both predictors accept data from source 1; then for t = 201; 202; :::; 400 predictor 1 specializes in source 2 and for t = 401; 402; :::; 600 predictor 2 specializes in source 1; after this time specialization is retained and no data are misallocated. This is a perfectly acceptable situation and corresponds to the case $X(t) \to 1$.

B. Experiment group B is presented in Figure 4. The setup is the same but now three sources are used; the third source is a double logistic time series.

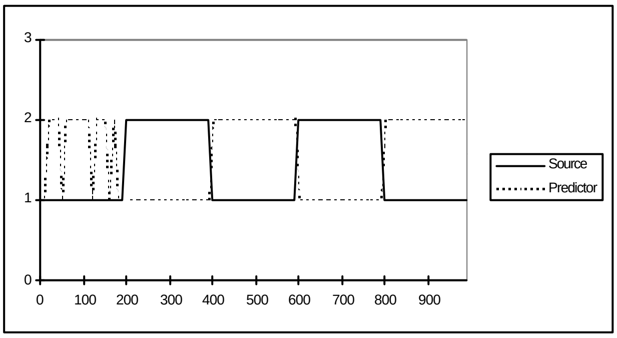

Figure 2: Classification time series for a logistic/ tent-map time series identification task. Solid line is source process Z(t) and dotted line is predictor process $\hat{z}(t)$:

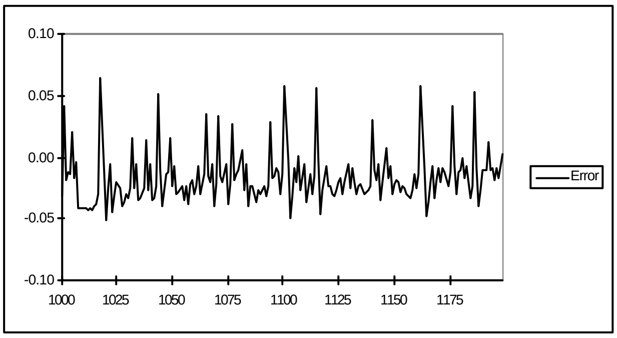

Figure 3: Prediction error time series for a logistic/ tent-map time series identification task.

6 Conclusions and Future Research

In this paper we have presented sufficient conditions for the convergence of a class of unsupervised, on-line, multiple model schemes for switching dynamical system identification and connected these conditions to PAC learnability. One model is trained for every active sub-system; allocation of training data to models is unsupervised, i.e. labeled data are not available. Observed data may be allocated to models according to their predictive performance, or more general data allocation criteria may be used.

Our analysis focuses on the data allocation problem, assuming that, given correctly allocated data, model training is not particularly hard. Hence, the main question discussed here is whether correct data allocation will be achieved. The answer, is given by Theorems 1 and 2: data allocation is succesful, provided that conditions B1, A1 and A2 are satisfied. These conditions are quite general and do not depend on particular properties of the systems, models or training algorithms used. Our analysis indicates that, for a problem which involves sub-systems with fairly distinct behavior and accurate models of these subsystems, a data allocation

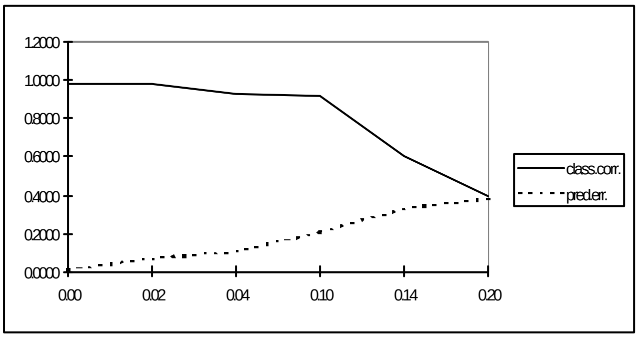

Figure 4: Result of Experiment Group B. Classification figure of merit c and prediction error d. Solid line corrsponds to c; dotted line corresponds to d. The horizontal axis indicates noise level A.

scheme will converge and specialization will take place. Conditions A1 and A2 can also be expressed in terms of PAC learnability for the case of predictive accuracy data allocation criterion. These conclusions are corroborated from numerical experiments which we conducted using a simple predictive multi-model system identification scheme.

The analysis presented here is limited to serial data allocation schemes; it is worthwhile investigating the behavior of parallel data allocation schemes (where all models compete simultaneously for obtaining an observed datum). Also, conditions A1 and A2 may be related to more specific convergence criteria by the introduction of capacity / complexity concepts[10].

References

[1] S. Chen, S.A. Billings and P.M. Grant, "Nonlinear system identification using neural networks", Int. J. of Control, 1990, vol.51, pp.1191-1214.

[2] A. Kehagias and V. Petridis, "Predictive modular neural networks for time series classification", Neural Networks, 1996, vol.10, pp.31-49.

[3] A. Kehagias and V. Petridis, "Time Series Segmentation using Predictive Modular Neural Networks", Neural Computation, 1997, vol.9, pp.1691-1710.

[4] V. Petridis and A. Kehagias, "A recurrent network implementation of time series classification", Neural Computation, 1996, vol.8, pp.357-372.

[5] V. Petridis and A. Kehagias. "Modular Neural Networks for MAP Classification of Time Series and the Partition Algorithm", IEEE Trans. on Neural Networks, 1996, vol.7, pp.73-86.

[6] V. Petridis and A. Kehagias. Predictive Modular Neural Networks: Time Series Applications. Amsterdam: Kluwer, 1998.

[7] T. Chen and H. Chen. "Universal Approximation to Nonlinear Operators by Neural Networks with Arbitrary Activation Functions and Its Application to Dynamical Systems", IEEE Trans. on Neural Networks, 1995, vol.6, pp.911-917.

[8] R. C. Williamson and U. Helmke, "Existence and Uniqueness Results for Neural Network Approximation", IEEE Trans. on Neural Networks, 1995, vol.6, pp.2-13.

[9] S. B. Holden and P. J. W. Rayner, "Generalization and PAC Learning: Some New Results for the Class of Generalized Single-Layer Networks", IEEE Trans. on Neural Networks, 1995, vol.6, pp.368-380.

[10] C. Mazza, "On the Storage capacity of Nonlinear Neural Networks", Neural Networks, 1997, vol.10, pp. 593-597.