Unsupervised Time Series Segmentation by Predictive Modular Neural Networks

Vas. Petridis and Ath. Kehagias

Dept. of Electrical and Computer Eng.,
Aristotle Univ. of Thessaloniki, Greece

Abstract

Consider a switching time series, produced by several randomly activated sources. The separation of incoming data into distinct classes may be effected using predictive modular neural networks, where each module is trained on data from a particular source. We present a mathematical analysis regarding the convergence of a quite general class of competitive, winner-take-all schemes which allocate data into classes, one class corresponding to each active source.

1 Introduction

Consider a time series generated by several randomly activated sources. Time series segmentation involves finding the active source at every time step. This has been examined in [3] (using local experts [2]) and in [1], where the following method is used. The observed time series is used as input to a bank of neural network neural predictive modules; at every time step the new observation y_t is allocated to the neural predictive module which yields minimum prediction error; then each module is retrained on the data so far allocated to it. In this manner each neural neural predictive module may be associated with a particular source, exhibiting minimum prediction error when this source is activated; hence, at every time step the active source is identified by the neural predictive module which has minimum error. In order to train each module, labeled data from each source must be available. If training must take place concurrently with segmentation, using the unlabeled measurements of the time series, then accurate segmentation requires on-line, unsupervised data allocation to the neural predictive modules. In this paper we examine the properties of a general class of competitive data allocation schemes which utilize predictive modular neural networks.[1]

2 Parallel Data Allocation

The source time series $Z(t); t = 1; 2; :::$, takes values in a finite source set $\pounds = f1; 2; :::; Kg$; the observation time series $Y(t); t = 1; 2; :::$, takes values in $<$

[1] To appear in \8th INTERNATIONAL CONFERENCE ON ARTIFICIAL NEURAL NETWORKS" September 1-4, 1998, Skoevde, Sweden.

(the set of real numbers). $Y(t)$ is generated by a function $Y(t) = F_{Z(t)}(Y(t-1); Y(t-2); \ldots; Y(t-L))$; where $F_1(\cdot); F_2(\cdot), \ldots, F_K(\cdot)$ are functions from $<^L$ to $<$. Hence, $Y(t)$ is determined by past observations and the current source. The time series segmentation consists in producing $\hat{Z}(t)$; an estimate of $Z(t)$, for times $t = 1; 2; \ldots$, which is equivalent to finding the source which is active for $t = 1; 2; \ldots$. Using K neural predictive modules of the form $\hat{Y}_k(t) = \hat{F}_k(Y(t-1); Y(t-2); \ldots; Y(t-L))$ we can compute the prediction $\hat{Y}_k(t)$ (for $k = 1; \ldots; K$) and set $\hat{Z}(t) = \arg\min_{k=1;2;\ldots;K} |Y(t) - \hat{Y}_k(t)|$; i.e. $Y(t)$ is allocated to the neural network of minimum prediction error. The problem can be decomposed into two subproblems: data allocation and predictor training; here we will deal with the former since, with accurately allocated data, the neural predictive module training subproblem can be solved using a variety of training algorithms. The following classification / training algorithm (a variation of which appears in [1]) is used.

At $t = 0$ K predictors are randomly initialized
For $t =$ 1,2,...
Observe $Y(t)$.
For $k = 1; 2; \ldots; K$ compute $\hat{Y}_k(t)$ and $|Y(t) - \hat{Y}_k(t)|$.
Assign $Y(t)$ to predictor nr. $\hat{Z}(t)$
(where $\hat{Z}(t) = \arg\min_{k=1;2;\ldots;K} |Y(t) - \hat{Y}_k(t)|$).
Retrain each predictive module on all data assigned to it.
Next t

Data allocation is performed in a competitive, winner-take-all manner, to the predictor of minimum error. The question discussed in this paper is whether (a) each neural predictive module will specialize in one source, accepting all or most data generated by this source and rejecting data from other sources, or (b) a neural predictive module will obtain data from more than one sources.

3 Convergence

Two sources. If two sources are active, the source process $Z(t)$ takes values in $\{1; 2\}$; at time t we have $\Pr(Z(t) = i) = \alpha_i$, $i = 1; 2$: Obviously $\alpha_1 + \alpha_2 = 1$; it is also assumed that: for $i = 1; 2$ we have $0 < \alpha_i < 1$: The observation $Y(t)$ is given by $Y(t) = F_{Z(t)}(Y(t-1); Y(t-2), \ldots, Y(t-L))$: For two neural predictive modules ($i = 1; 2$) we have $\hat{Y}_i(t) = \hat{F}_i(Y(t-1); Y(t-2), \ldots, Y(t-L))$. The allocation process $W(t)$ takes values in $\{1; 2\}$; $W(t) = i$ means that $Y(t)$ is allocated to the i-th neural predictor. If, at time t, y_t is generated by source i and allocated to neural predictor j, then $Z(t) = i$ and $W(t) = j$. The processes $M_{ij}(t)$ ($t = 1; 2; \ldots, i; j = 1; 2$) are defined by

$$M_{ij}(t) = \begin{cases} 1 & \text{if } Z(t) = i; W(t) = j \\ 0 & \text{else;} \end{cases}$$

and the processes $N_{ij}(t)$ (where $t = 1; 2; :::$ and $i; j = 1; 2$) are de¯ned by $N_{ij}(t) = \sum_{s=1}^{t} M_{ij}(s)$. Hence $N_{ij}(t)$ indicates the total number of source i samples assigned to neural predictive module j, up to time t. The variable $X(t)$; denotes the total specialization of the system:

$$X(t) = [N_{11}(t) - N_{21}(t)] + [N_{22}(t) - N_{12}(t)]:$$

The data assignment probabilities (for neural predictive module 1) depend on $X(t)$. In case $X(t)$ is large and positive, at least one of $[N_{11}(t) - N_{21}(t)]$ and /or $[N_{22}(t) - N_{12}(t)]$ must be large and positive, which means that either neural predictive module nr.1 has received a large surplus of source nr.1-generated data, or neural predictive module nr.2 has received a large surplus of source nr.2-generated data, or both. Similar remarks hold in case $X(t)$ is large and negative. Hence, it is reasonable to assume that the data assignment probabilities (for neural predictive module 1) depend on $X(t)$:

$$f(n) = Pr(W(t) = 1 | Z(t) = 1; X(t - 1) = n);$$
$$g(n) = Pr(W(t) = 1 | Z(t) = 2; X(t - 1) = n):$$

In other words, $f(n)$ is the probability that neural predictive module 1 accepts a datum from source 1, given that so far it has accepted n more data from source 1 than from source 2, while $g(n)$ is the probability that neural predictive module 1 accepts a datum from source 2, given that so far it has accepted n more data from source 1 than from source 2. Regarding these probabilities, the following assumptions are made.

A1 For $n = :::; -1; 0; 1; :::$ $f(n) > 0$; $\lim_{n \to -1} f(n) = 0$; $\lim_{n \to +1} f(n) = 1$;

A2 For $n = :::; -1; 0; 1; :::$ $g(n) > 0$; $\lim_{n \to -1} g(n) = 1$; $\lim_{n \to +1} g(n) = 0$:

By assumption A1, if neural predictive module 1 has accumulated many more data from source 1 than from source 2, then it will be very likely to accept an additional datum generated from this source and will be very unlikely to accept an additional datum from source 2. This is reasonable: if the neural predictive module has been trained on data mostly originating from source nr. 1, rather than from nr. 2, than it will exhibit improved performance on source nr.1 data and deteriorated performance on source 2 data. Similar remarks can be made regarding assumption A2. It must be stressed that A1 and A2 refer to the combination of time series, network architecture, training law and data allocation algorithm. It is not necessary to take into account particular characteristics of any of the above components; it is only required that A1 and A2 hold true, which may be the case for various combinations of time series, network architecture, training law and data allocation algorithm.

The data allocation procedure described above, implies that $X(t)$ is Markovian. The transition probabilities (for $m; n = 0; \pm 1; \pm 2; :::$) can be obtained from the data allocation method and are $p_{n;m} = 0$ if $|n - m| \neq 1$, $p_{n;n-1} = \frac{1}{2} \cdot g(n) + \frac{1}{4} \cdot (1 - f(n))$, $p_{n;n+1} = \frac{1}{4} \cdot f(n) + \frac{1}{2} \cdot (1 - g(n))$. Hence, convergence can be studied using methods from the theory of Markov chains. We have established two convergence theorems; in this paper we omit the proofs because of space limitations. The ¯rst theorem ensures convergence of $X(t)$.

Theorem 1 If conditions A1, A2 hold, then

$$\text{(i)} \quad \Pr\left[\lim_{t \to \infty} |X(t)| = +\infty\right] = 1;$$

$$\text{(ii)} \quad \Pr\left[\lim_{t \to \infty} X(t) = +1\right] + \Pr\left[\lim_{t \to \infty} X(t) = -1\right] = 1.$$

From (i) it is seen that total specialization goes to infinity; from (ii) it is seen that at least one neural predictive module will (in the long run) accumulate either a lot more source nr.1 samples than source nr.2 samples ($X(t) \to +1$) or a lot more source 2 samples than source 1 samples ($X(t) \to -1$). The total probability that one of these two events will take place is one, i.e. one neural predictive module will certainly specialize in one of the two sources.

Theorem 2 If conditions A1, A2 hold, then

$$\Pr\left[\lim_{t \to \infty} \frac{N_{21}(t)}{N_{11}(t)} = 0 \,\middle|\, \lim_{t \to \infty} X(t) = +1\right] = 1;$$

$$\Pr\left[\lim_{t \to \infty} \frac{N_{12}(t)}{N_{22}(t)} = 0 \,\middle|\, \lim_{t \to \infty} X(t) = +1\right] = 1;$$

$$\Pr\left[\lim_{t \to \infty} \frac{N_{11}(t)}{N_{21}(t)} = 0 \,\middle|\, \lim_{t \to \infty} X(t) = -1\right] = 1;$$

$$\Pr\left[\lim_{t \to \infty} \frac{N_{22}(t)}{N_{12}(t)} = 0 \,\middle|\, \lim_{t \to \infty} X(t) = -1\right] = 1.$$

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

Only a very brief sketch will be of the proofs of the above theorems is given. Regarding Theorem 1, it is proved that the Markovian process $X(t)$ is transient; i.e. that w.p.1 (with probability one) $X(t)$ will spend only a finite amount of time in any particular state. Then it follows that $|X(t)|$ must go to infinity w.p.1, which is (i); (ii) follows easily. Regarding Theorem 2, we exploit the fact that, if $X(t)$ goes to 1, then transitions to lower states are highly improbable. Such transitions are "counted" by the process $N_{21}(t)$. This process is dependent, but it can be compared to an auxiliary process $\overline{N}_{21}(t)$, which has a larger probability of transitions to lower states and is independent;

in fact it is a sequence of Bernoulli trials, and its properties are easily obtained. By appropriate construction of $\overline{N}_{21}(t)$ it can be proved that $\overline{N}_{21}(t)=t$ goes to zero with probability one; then relating $\overline{N}_{21}(t)$ and $N_{21}(t)$ it can be shown that also $\overline{N}_{21}(t)=t$ goes to zero with probability one. A similar argument, depending on an uxiliary process $\overline{N}_{11}(t)$ is used to show that $N_{11}(t)=t$ goes to ¼1. Then the ¯rst conclusion of the theorem follows easily. The remaining conclusions are proved similarly.

Many Sources. The case of more than two sources is treated here by an informal argument; a more formal presentation in terms of convergence theorems will be reported in the future. Consider the case of K sources ($K > 2$) and a data allocation scheme starting with two neural predictors, and adding more predictors \as needed" (for instance whenever the prediction error exceeds a certain threshold). Consider two sources: source 1 and composite source $[2; 3; :::; K]$. By Theorems 1 and 2, in the long run one neural predictive module will mostly receive data from one source. In the long run the second neural predictive module will mostly receive data from the other source. Hence the data are separated into two sets: those generated by source 1 and those generated by all other sources. Reapplying the data allocation scheme on the composite data set, we will obtain two new neural predictive modules, with one specializing in source 2 data and the other specializing in sources 3, 4, ... , K. After suf¯cient time has elapsed, neural predictive module 2 will specialize in source 2, while neural predictive module 3 will mostly receive data from sources 3,4, ..., K. The same argument can be repeated for sources 3, 4, ..., K, adding neural predictive modules as needed, resulting in one neural predictive module specializing in each source.

4 Experiments

Exp. Group A. Four sources have been used: (a) for $Z(t) = 1$, a logistic time series of the form $y(t) = f_1(y(t ¡ 1))$, where $f_1(x) = 4x(1 ¡ x)$; (b) for $Z(t) = 2$, a tent-map time series of the form $y(t) = f_2(y(t ¡ 1))$, where $f_2(x) = 2x$ if $x \in [0; 0:5)$ and $f_2(x) = 2(1 ¡ x)$ if $x \in [0:5; 1]$; (c) for $Z(t) = 3$, a double logistic time series of the form $y(t) = f_3(y(t ¡ 1)) = f_1(f_1(y(t ¡ 1)))$ and (d) for $Z(t) = 4$, a double tent-map time series of the form $y(t) = f_4(y(t ¡ 1)) = f_2(f_2y(t ¡ 1)))$. The four sources are activated consecutively, each for 100 time steps, giving an overall period of 400 time steps. Ten such periods are used, resulting in a 4000-steps time series. The task is to discover the four sources and the switching schedule by which they are activated. At every step y_t is mixed with additive white noise uniformly distributed in the interval $[¡ A=2; A=2]$. The neural predictive modules used are 1-5-1 sigmoid neural networks. In every experiment performed, all four sources are eventually identi¯ed. This takes place at some time T_c, which is di®erent for every experiment. After time T_c, a classi¯cation ¯gure of merit is computed. It is denoted by $c = T_2=T_1$, where T_1 is the total number of time steps after T_c, and T_2 is the number of correctly classi¯ed time steps after T_c. Table 1 shows the results of the experiments.

Exp. Group B. Here we consider a time series obtained from three sources of the Mackey-Glass type. The time series evolves in continuous time and satisfies the differential equation : $\frac{dy}{dt} = -0.1y(t) + \frac{0.2y(t-t_d)}{1+y(t-t_d)^{10}}$: For each source a different value of the delay parameter t_d was used, namely t_d= 17, 23 and 30. The time series is sampled in discrete time, at a sampling rate Δ = 6, with the three sources being activated alternately, for 100 time steps each. The final result is a time series with a switching period of 300 and a total length of 4000 time steps. The time series is observed at various levels of additive observation noise; results expressed in terms of the parameters c and T_c appear in Table 1. Segmentation is again quite accurate for fairly high noise levels.

Table 1: Segmentation Results for Experiment Groups A and B

Exp. Group A	A	0.00	0.05	0.10	0.15	0.20
	T_c	500	1800	1800	800	2200
	c	0.982	0.969	0.947	0.529	0.529
Exp. Group B	A	0.00	0.05	0.10	0.15	0.20
	T_c	1700	1100	1300	3500	1200
	c	0.978	0.977	0.853	0.935	0.664

5 Conclusion

In this paper we have presented two theorems (regarding the convergence of competitive data allocation) which state that, if the general conditions A1 and A2 are satisfied, data allocation will result in succesful predictor specialization. The competitive data allocation method may also be called \parallel", in contradistinction to a \serial" method, where a threshold is fixed and the first predictive module with prediction error below the threshold receives the new incoming datum (at every time step the predictive modules are considered in a specified order).We have performed a convergence analysis of the serial data allocation case, which yields results similar to the parallel case; this analysis is presented elsewhere. Hence we now have general conditions which ensure convergence of the data allocation scheme for both the serial and parallel case.

References

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

[2] M.I. Jordan and R.A. Jacobs. 1994. \Hierarchical Mixtures of Experts and the EM Algorithm", Neural Computation, vol.6, pp. 181-214.

[3] K. Pawelzik, J. Kohlmorgen and K.R. Muller. 1996. \Annealed Competition of Experts for a Segmentation and Classification of Switching Dynamics", Neural Computation, vol.8, pp.340-356.