Classification of Time Series Data with Nonlinear Similarity Measures

Thomas Schreiber and Andreas Schmitz
Physics Department, University of Wuppertal, D-42097 Wuppertal, Germany

(Received 13 January 1997)

We address the problem of unsupervised classification of time series recordings. The aim is to form groups of sequences or segments of a longer sequence which show similar dynamical behavior. Several measures of similarity between two time series are considered. Groups or clusters are then formed by minimizing a suitable cost function using simulated annealing. Apart from the classification of systems, the method can be applied to the monitoring of abrupt or continuous changes in nonstationary signals. We further discuss the inclusion of supervision and propose the use of similarity measures in tests for nonlinearity based on surrogate data. [S0031-9007(97)03799-X]

PACS numbers: 05.45.+b

In many areas of science and engineering, systems can be studied through the characteristic time evolution of observable properties. Different types of systems, or different states of a single system, can then be distinguished by analyzing appropriate time sequences. Electrocardiographic recordings, for example, are routinely classified by the cardiologist who looks for typical patterns associated with certain diseases. Malfunction in mechanical equipment can be detected in a time series by comparing the system's dynamics during different periods of operation. Usually, time series classification is done by computing some characteristic parameter (or a small number of such parameters) for each time series in question. Then the actual classification is done on the base of this table of parameters. Often, this has to be done in an *unsupervised* fashion since it is not known *a priori* which measured parameters correspond to which class. In the case of a scalar characteristic γ, γ has to follow a K-humped distribution in order to distinguish K different classes.

In this paper we propose a different approach which avoids the severe loss of information which occurs when a time series is summarized by one or a few characteristic numbers. Rather than comparing such numbers, we will compare the time series themselves. Such an approach has been taken in the context of stationarity testing in Refs. [1–3], where it has been demonstrated that similarity measures like cross predictions contain important additional information as compared to the corresponding measures for single time series. Here we show how to use the knowledge about similarities of time series to split the set of sequences into meaningful groups, or *clusters*. With the usual reduction to a single characteristic number, this separation has to be done in the one-dimensional space spanned by this quantity. A set of mutual similarities allows us to work in an abstract space of dynamical properties without having to specify a base or even its dimension. A similar approach leads to promising results in the special context of the classification of the morphology of electroencephalographic recordings [1].

Our purpose is to partition a collection of J time sequences into K groups (or clusters) of series with similar dynamics. Of course, the sequences can also be segments of one long recording. There is rich literature on the general classification problem (see, e.g., Ref. [4]). We are here interested in the case that the objects (these are in our case time sequences) have to be grouped on the base of a table of *dissimilarities* only. Proper coordinates in a metric space are not available [5]. We cannot expect that the dissimilarities γ_{ij} between series i and j are proper distances fulfilling the triangle inequality. They will however be taken to be (or made) symmetric. For "similar" objects, γ_{ij} should be smaller than for "dissimilar" objects. In this paper we will use cross-prediction errors (Ref. [2]) or normalized cross-correlation sums (Ref. [6]) as dissimilarities. Note that two time series realizations respecting the same dynamics cannot be expected to have $\gamma_{ij} = 0$ *exactly*.

In order to classify the J objects into K clusters let us define a membership index $u_i^{(\nu)}$ which is 1 if object i is in cluster ν and 0 otherwise [7]. We have $\sum_{\nu=1}^K u_i^{(\nu)} = 1$ and a cluster is defined as $C^{(\nu)} = \{i\colon u_i^{(\nu)} = 1\}$. We want to form clusters such that the average dissimilarity of objects *within* each cluster is minimal. If we define $|C^{(\nu)}| = \sum_{i=1}^J u_i^{(\nu)}$ to be the number of objects in cluster ν, then the average dissimilarity of object i to the objects in cluster ν can be written as

$$D_i^{(\nu)} = \frac{1}{|C^{(\nu)}|} \sum_{j=1}^{J} u_j^{(\nu)} \gamma_{ij} \,. \qquad (1)$$

The average dissimilarity of objects within cluster ν is then $D^{(\nu)} = (1/|C^{(\nu)}|) \sum_{i=1}^J u_i^{(\nu)} D_i^{(\nu)}$. For the total average dissimilarity we take a sum over all clusters, weighted by their numbers of members, $D = (1/K) \times \sum_{\nu=1}^K |C^{(\nu)}| D^{(\nu)}$. Thus an optimal partitioning into clusters can be found by minimizing the cost function

$$E = KD = \sum_{\nu=1}^{K} \frac{1}{|C^{(\nu)}|} \sum_{i,j=1}^{J} u_i^{(\nu)} u_j^{(\nu)} \gamma_{ij} \,. \qquad (2)$$

In this motivation, some arbitrary choices have been made. In particular we could have weighted the clusters with higher powers of their size, thereby introducing a bias towards clusters with equal size. Other alternatives are discussed in the literature [4].

We numerically minimize E in Eq. (2) by simulated annealing, which is particularly useful for discrete minimization problems with false minima. The problem is formulated as a thermodynamic system. The temperature T is decreased slowly to allow the system to settle down to its ground state. In this analogy the cost function of a configuration plays the role of the energy of a microstate. The goal is to reach the absolute minimum, the ground state at $T = 0$. In a Metropolis simulation, a new configuration is generated by moving an object from one cluster to another. The new configuration is accepted with a probability given by the Boltzmann probability distribution $p = \exp(-\Delta E/T)$ if $\Delta E > 0$ and $p = 1$ if $\Delta E \leq 0$. For a discussion of simulated annealing, see, e.g., Ref. [8]. In all examples where the global minimum was known, the annealing scheme converged fast and reliably to that configuration. Of course, since the dissimilarities are subject to fluctuations, the minimum of E may not always be given by the desired clustering.

For time series objects, many dissimilarity measures can be thought of. Since we want to assign a high degree of similarity to two different realizations of the same process, although they might be out of phase with each other, we will quantify the closeness of dynamical systems or their attractors rather than individual time series. The question of similarity between the *dynamics* of two signals has been addressed previously in Refs. [1–3,6]. The question if two *trajectories* are related by a mapping of certain properties is discussed in Refs. [9,10].

If the time series in question can be adequately described by finite order linear models then it seems most plausible to define similarity through the coefficients of this model. In this paper we mainly have applications in nonlinear settings in mind. Most nonlinear discriminating statistics which are commonly used are straightforwardly generalized to similarity measures. Let $X = \{x_k\}$ and $Y = \{y_l\}$, $k, l = 1, \ldots, L$ be two time series we want to compute the dissimilarity of. First we form delay vectors as usual, $\vec{x}_k = (x_{k-(m-1)\tau}, x_{k-(m-2)\tau}, \ldots, x_k)$. Following Ref. [6], the Grassberger-Procaccia correlation sum [11] can then be generalized to

$$C_{XY}(\epsilon) = C_0 \sum_{k=k_0}^{L} \sum_{l=l_0}^{L} \Theta(\epsilon - \|\vec{x}_k - \vec{y}_l\|), \qquad (3)$$

where C_{XX} is the usual correlation sum (apart from the fact that entries with small $|k - l|$ have to be excluded from the sum), $k_0 = l_0 = (m - 1)\tau + 1$, and C_0 is a normalization constant. We can take either C_{XY} directly, or $C_{XY}/\sqrt{C_{XX}C_{YY}}$, or the square of any of these as a similarity measure. The cross-correlation sum (3) has been used as a similarity measure for the study of nonstationary time series in Ref. [3].

One way to estimate the maximal Lyapunov exponent from time series data is by observing the divergence of close-by pairs of trajectories [12]. Here the generalization could be to study pairs with one member from time series i and the other from series j. These are however not expected to diverge exponentially whence the divergence rate will become scale dependent.

For prediction errors a generalization has been described in Ref. [2]. Nonlinear predictors F_X and F_Y are formed using X and Y, respectively, as databases. Different parametric models F_X and F_Y could in principle be distinguished by the difference in the parameters. The relation between the parameters and the dynamical behavior of F, however, is usually quite involved; models with different parameters may be almost equivalent on the data. Therefore we compute the (nonsymmetric) cross-prediction error $\sigma_{X,Y}$, defined as the root mean squared error of F_X applied to Y:

$$\sigma_{X,Y}^2 = \frac{1}{L'} \sum_{k=(m-1)\tau+1}^{L-1} \|\vec{y}_{k+1} - F_X(\vec{y}_k)\|^2,$$

where $L' = L - (m - 1)\tau - 1$ is the number of delay vectors available. In principle, F can be taken from any class of time series models. We prefer nonparametric models like the locally constant or locally linear approximations. The former are known in the statistical literature as *kernel estimators*. In this paper we use the same locally constant predictor as in Ref. [2]. An alternative way to compare two predictive models is through the average difference in the predictions they yield; see Ref. [1].

Let us illustrate the use of dissimilarities and cluster algorithms with a couple of example time series. If the series have been generated in groups of similar dynamics, successful clustering can be simply stated by comparison to the correct grouping. Once the clusters have been formed, they can also be used to construct a coordinate space for time series objects i. The coordinates are simply the average dissimilarities $D_i^{(\nu)}$ of object i to the objects in cluster ν, as given by Eq. (1). This representation will enable us to judge if there are distinct clusters or if the sequences differ through a continuously changing parameter.

First, consider a generalized baker map

$$\begin{array}{lll} v_n \leq \alpha\colon & u_{n+1} = \beta u_n, & v_{n+1} = v_n/\alpha\,, \\ v_n > \alpha\colon & u_{n+1} = 0.5 + \beta u_n, & v_{n+1} = \frac{v_n - \alpha}{1-\alpha}\,, \end{array}$$

with $\alpha = 0.4$. The parameter β can be varied without changing the positive Lyapunov exponent [13]. We create 50 sequences with $\beta = 0.6$ and 50 with $\beta = 0.8$, each of them of length 400. We calculate cross-prediction errors $\sigma_{i,j}$ and from these form a symmetric dissimilarity matrix: $\gamma_{ij} = (\sigma_{i,j} + \sigma_{j,i})^2/4\sigma_{i,i}\sigma_{j,j}$. This particular choice was made in order to demonstrate that the information contained in the diagonal entries $\sigma_{i,i}$ is not used ($\gamma_{ii} = 1 \forall i$). Two clusters are formed by minimizing E. In Fig. 1, $D_i^{(2)}$ is plotted against $D_i^{(1)}$.

Two distinct groups can easily be seen and the algorithm forms exactly the two desired clusters.

The second example is also a baker map time series, but now β is changing continuously from 0 to 1. It is split into 20 segments of length 2000 and between them cross-correlation sums [6] C_{ij} are calculated. As a dissimilarity measure we choose $\gamma_{ij} = |1 - C_{ij}/\sqrt{C_{ii}C_{jj}}|$ which is symmetric in i and j. In Fig. 2, distances $D_i^{(\nu)}$ are plotted for two clusters $\nu = 1, 2$ and in Fig. 3 for three clusters, $\nu = 1, 2, 3$. From these two plots we can see that there is a continuous change of parameter; instead of a sharp "jump" in the distances they form a continuous path.

As a third example we consider an ensemble of m uncoupled Ulam maps, $x_{n+1,i} = 2 - x_{n,i}^2$, $i = 1, m$. For each $m = 1, \ldots, M$ we create ten time series each of length 1000 by recording $s_{n,m} = \sum_{i=1}^{m} x_{n,i}$ and normalizing to unit variance. Cross-prediction errors with embedding dimension $m = 3$ and cluster analysis are used in order to distinguish how many maps have been summed together. Up to $M = 19$, the algorithm separates the groups without mistakes. At $M = 20$ the first 15 faults (out of 200 segments) occur. The algorithm was able to distinguish two groups of ten series each with $m = M - 1$ and $m = M$ maps up to $M = 30$ without error.

Let us finally show how the concept of classification based on dissimilarities can be used in a surrogate data test for nonlinearity [14]. The idea of such a test is to create a Monte Carlo sample of "surrogate" data sets which are constrained to have the same linear properties as the data to be tested but are otherwise random. Then a nonlinear statistic γ is computed on the data and the surrogates and a statistical test is performed to decide if γ_0 of the data is significantly different from the distribution of γ_i measured on the surrogates. In other words, the total set of J sequences containing data and surrogates is classified into two groups one of which has just one member. In the absence of nonlinear structure, this one-element cluster will contain the original data by chance with probability $1/J$. Thus, if we find that the classification singled out the original data, we can reject the null hypothesis at the $(1 - 1/J) \times 100\%$ level of significance. It is obvious that this scheme can be modified by the use of dissimilarities γ_{ij}. In order to check how this modification affects the discrimination power of the test, we perform 600 tests with data sets of length 1000 from an NMR laser experiment [15] which have been contaminated by 80% additive in-band noise. Each test is done at the 90% level of significance, that is, using nine surrogate data sets, created according to Ref. [16]. Cross-prediction errors are used as above. If the classification is done using diagonal terms γ_{ii} only (this corresponds to the usual surrogate data test), nonlinearity is correctly detected in $61.5\% \pm 2.5\%$ of the cases. Using the full matrix γ_{ij} and forming two clusters of size 1 and 9, respectively, correctly singled out the data in $69\% \pm 2\%$ of all trials. Thus the discrimination power is significantly higher.

The usefulness of similarity measures from nonlinear dynamical systems theory has been recently pointed out by several authors [1–3]. We discussed different propositions for dissimilarity measures, in particular, cross-prediction errors [2] and cross-correlation integrals [6]. We demonstrated how a table of dissimilarities can be used for the unsupervised classification of time series with the help of cluster algorithms. In the scheme adopted in this paper, learning can be supervised to any desired extent by fixing the cluster memberships of any number of objects which then constitute the training set. These memberships are simply excluded from the minimization procedure. We showed that unsupervised classification can also be used to increase the power of Monte Carlo tests for nonlinearity.

A desirable generalization of the approach taken in this paper is to allow a smooth change of coordinates

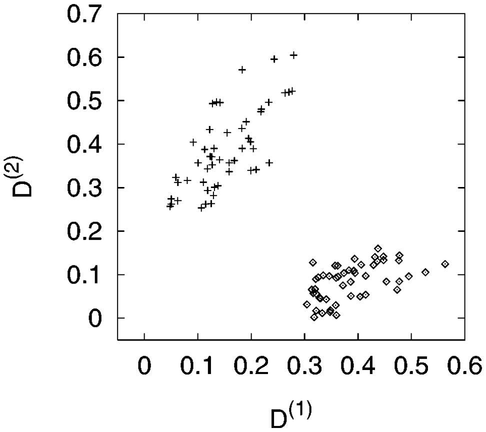

FIG. 1. Distances of objects from two clusters generated for 100 time series in two groups of 50. Baker map with $\beta = 0.6$ in the first group and $\beta = 0.8$ in the second group.

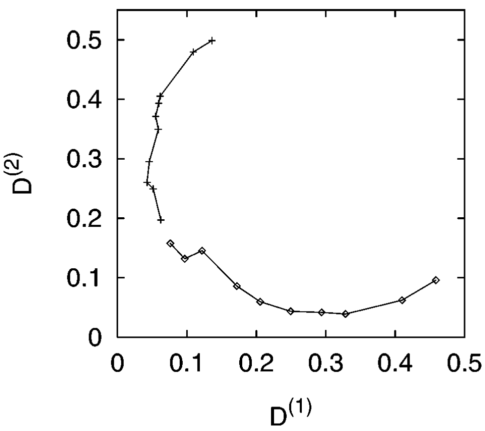

FIG. 2. Distances $D_i^{(\nu)}$ of objects from two clusters generated for a time series split into 20 segments. The membership to the two clusters is denoted by different symbols. Baker map with β varying continuously from 0 to 1.

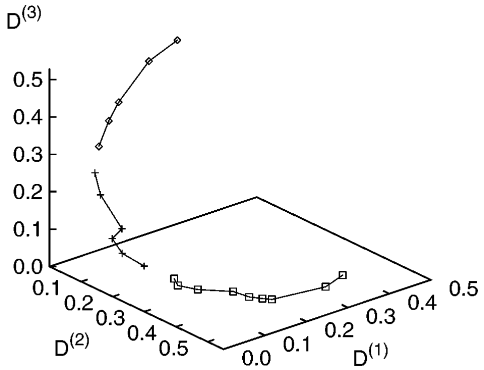

FIG. 3. Same as Fig. 2, except that now three clusters are formed.

between dynamical systems considered similar. However, the construction of a reliable measure of dissimilarity with that property is still an open problem and subject to ongoing research.

We thank Peter Grassberger, Michel Le van Quyen, and Francisco Varela for useful discussions; in particular, we thank Johannes Müller-Gerking for drawing our attention to Ref. [1]. This work was supported by the SFB 237 of the Deutsche Forschungsgemeinschaft.

[1] J.L. Hernández, R. Biscay, J.C. Jimenez, P. Valdes, and R. Grave de Peralta, Int. J. Bio-Med. Comput. **38**, 121 (1995).

[2] T. Schreiber, Phys. Rev. Lett. **78**, 843 (1997).

[3] R. Manuca and R. Savit, Physica (Amsterdam) **99D**, 134 (1996).

[4] L. Kaufman, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York, 1990).

[5] A slightly different approach not pursued in this paper would describe each time series object by a vector of discriminating quantities. Then the formation of groups could be done in the space of these vectors. The main difficulty is to find several meaningful but nonredundant quantities that span an appropriate space.

[6] H. Kantz, Phys. Rev. E **49**, 5091 (1994).

[7] Alternatively, one could form “soft” or “fuzzy” clusters by taking $u_i^{(\nu)} \in [0, 1]$ as a probability to find object i in cluster ν.

[8] W.H. Press *et al., Numerical Recipes* (Cambridge University Press, Cambridge, England, 1995), 2nd ed., Sect. 10.9.

[9] L.M. Pecora, T.L. Carroll, and J.F. Heagy, Phys. Rev. E **52**, 3420 (1995).

[10] S.J. Schiff, P. So, T. Chang, R.E. Burke, and T. Sauer, Phys. Rev. E **54**, 6708 (1996).

[11] P. Grassberger and I. Procaccia, Physica (Amsterdam) **9D**, 189 (1983).

[12] H. Kantz, Phys. Lett. A **185**, 77 (1994).

[13] H.G. Schuster, *Deterministic Chaos: An Introduction* (Physik Verlag, Weinheim, 1988).

[14] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J.D. Farmer, Physica (Amsterdam) **58D**, 77 (1992).

[15] M. Finardi, L. Flepp, J. Parisi, R. Holzner, R. Badii, and E. Brun, Phys. Rev. Lett. **68**, 2989 (1992).

[16] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).