QUANTIFYING SIMILARITY BETWEEN TIME SERIES USING THE SOM

Hui Liu, José C. Príncipe, J. Chris Sackellares*
Computational NeuroEngineering Laboratory (CNEL) University of Florida
Gainesville, Florida, USA
*Department of Neurology & Biomedical Engineering, University of Florida
Gainesville, Florida, USA
liuh, principe@cnel.ufl.edu
sackellares@mbi.ufl.edu

Abstract –*We propose a practical method to quantify the similarity between time series through its delay vector distribution. The method uses the Self-organizing map (SOM) to represent the time series in phase space and to build an histogram of the winner's processing elements from a known time series segment (that works as a template). The Kulback-Leibler (KL) divergence or the Correlation Coefficient (CC) of the transition matrix of winners is applied to estimate the similarity of the template with the one constructed on line from other window segments of the time series. The method performs at the same level as Diks' test, but it is computationally much simpler and can be run on line.*

Key words – SOM, Time Series Similarity

1 Introduction

Research in time series analysis has been employing the linear generation model for many years. Nonlinear dynamic theory has recently been introduced into time series analysis and lifted some of the difficulties faced by linear theory. In nonlinear dynamic theory the first step is to reconstruct the underlying multivariate dynamical system from the observed one-dimensional time series using Takens' embedding theorem [1] or alternative embedding methods [2]. Many researchers have documented phase space reconstruction either in the context of static geometry or state dynamical transitions.

Once the attractor is properly reconstructed, different alternatives exist to further study the dynamics of the generating system: (a) global attractor properties can be quantified by the correlation integral Grassberger & Procaccia [3] and its extensions Albano, Rapp and Passamante [4], Schreiber and Schmitz [5], and Kantz [6], or by Lyapunov exponents as proposed by Wolf et al.[7]; (b) nonlinear models can be estimated using dynamical neural network models8. Alternatively, one may be interested in quantifying similarity among attractors by estimating the delay vector density distribution as proposed by Wright and Schult [9], Wright [10], or Diks et al [11]. The method presented in this paper falls in this class of problems.

The Diks test detects differences between delay vectors distributions. It estimates the continuous delay vectors distributions using Parzen windowing with a Gaussian kernel and the difference between two sets of delay vectors distributions is then evaluated via a statistics

based on the estimated pdfs. The test can be easily adapted to quantify time series similarity. The major issue of Diks test is the prohibitive computation cost in high dimensions. The expense stems from the estimation of distances in probability spaces for continuous variables. There are efficient methods for density estimation overviewed in [12], the efficiency of which comes from discretization. The observation space is divided into non-overlapping hypercubes, and the discrete density distribution reduces to counting samples within each hypercube. Here we propose using the SOM to discretize the data distribution estimation process. The building blocks used in the proposed similarity quantification methodology are not new, however the way we propose to use them to evaluate similarity in time series is novel and very effective. The paper is composed of method description in section 2, simulation in section 3, and a short discussion in section 4. In the simulation section the method is tested on numerically generated data from the Mackey-Glass time series and it is compared with both Diks test (the continuous density estimation with higher computation cost) and the box counting method in [12].

2 Method

The centerpiece of our methodology is the utilization of the self-organizing map (SOM) as an infrastructure to model the trajectories in phase space with the added advantage of discretization and neighborhood preservation. The SOM output space is a discrete projection preserving neighborhoods of the high dimensional phase space. Therefore different trajectories in reconstruction space will be mapped into a different set of winning processing elements (PEs). The idea is to represent each trajectory by the histogram of the winning PEs, which will provide a signature in a discrete and finite space with which other histograms created from other trajectories can be compared against in a metric sense. We propose to utilize the Kullback-Leibler divergence to estimate the distance between the two histograms, with the added advantage that they always exist in the low dimensional output space of the SOM and the estimation can use discrete probabilities due to the discretization operated by the SOM. Alternatively, one can compare the transition matrices of the winning PEs for each trajectory using a metric resembling the correlation coefficient since the SOM preserves neighborhood relations. These operations will be explained next.

2.1 SOM template building

In this paper the SOM training algorithm used is taken from [13]. The stopping criterion is a preset value of the gradient of mean square error (MSE) between training samples and winning PEs weights from epoch to epoch. When the average gradient reached 0.001 in the past 100 epochs, the training stopped.

To train the SOM for a given scalar time series $x(i)$, for $i = 1, 2, \ldots, M$, where i is the time index, assume it is generated by a D dimensional dynamical system. From Takens' embedding theorem[1], the corresponding m-dimensional phase space ($m \geq 2D+1$) can be constructed from the time series $\boldsymbol{X_k} = [x(k), x(k+L), x(k+2L), \ldots, x(k+(m\text{-}1)L)]^T$, where L is the time delay. The delay vector sequence $\boldsymbol{X_k}$ $\{k = 1,2,\ldots, K,$ and $K = M\text{-}(m\text{-}1)L\}$ constitutes a trajectory in phase space. This phase space is mapped onto a two dimensional SOM represented by N ($N << K$) PEs each with weight $\boldsymbol{W_j}=[w_{j1}, w_{j2}, \ldots w_{jm}]^T$, $j = 1,2,\ldots N$.

Given a trained SOM, let us create a template corresponding to a segment of the time series. Calculate the winner sample distribution $\boldsymbol{Q}$ over the SOM as the ratio between the numbers of times the j^{th} PE is fired divided by the number of training samples. Calculate the transitions distribution between i^{th} and j^{th} PEs, $\boldsymbol{Y_{ij}}$ as the ratio of the sample transitions from i^{th} PE to j^{th} PE

divided by the total number of transitions (K-1). Either the SOM histogram $\boldsymbol{Q}$ or the transition matrix $\boldsymbol{Y}$ extracts a template for the known segment of the time series. Figure 1 shows the procedure to create these quantities diagrammatically. Figure 2 is an example of a 2D SOM trained with 2000 training samples of the Lorenz system (x variable, embedding dimension 3, delay 2). The template has 100 PEs. Figure 2 (a) plots in 2D space the winning PEs' for the first 500 training samples, while (b) shows the histogram of the PE winners also in 2D space. This histogram represents in the 2D space a projection of the density of points in the trajectory in the original phase space. Notice that it is discretized through the SOM, which simplifies the distance calculation latter, but also introduces a quantization error.

2.2 Similarity measure

After the template for the segment is constructed the goal is to find similar segments in the remaining portion of the time series. Two similarity measures are considered: the Kullback-Leibler divergence measuring the delay vector distribution similarity; and the Correlation Coefficient measuring a first order dynamic transition similarity.
The phase space of a window of the test data is created using the same embedding parameters of the training samples, and a histogram P of the wining PEs over the SOM is created. The similarity between the template and the test window histogram is computed as:

$$KL(P,Q) = \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{p_j}\right) \tag{1}$$

Where N is the number of PEs in the SOM template, Q is the 2D histogram of the template and P is the 2D histogram of the test window histogram.
As for the distance between transitions distributions, we define it as:

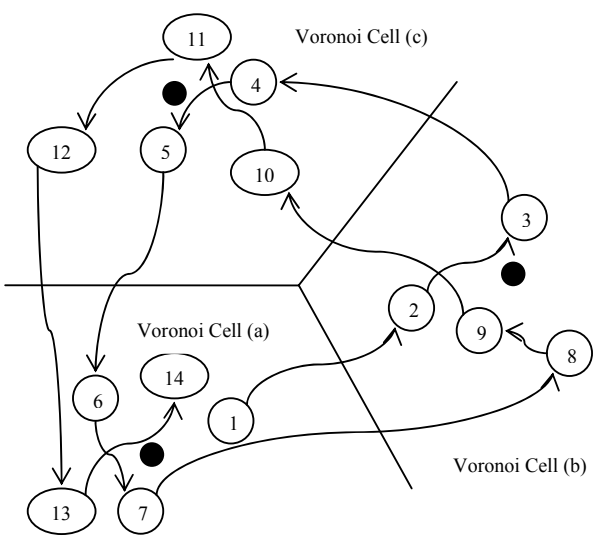
$$CC(Z,Y) = \frac{1}{N} * \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} Z_{ij} * Y_{ij}}{\sqrt{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} Z_{ij} * Z_{ij}\right) * \left(\sum_{i=1}^{N}\sum_{j=1}^{N} Y_{ij} * Y_{ij}\right)}} \tag{2}$$

Where N is the number of PEs in the SOM template, Y is the transition matrix of the template, and Z is the transition matrix of the test window, built in the same way as for the template. Effectively this distance is the correlation coefficient between the norms of the two transition matrices. It is based on the Cauchy-Schwartz distance presented in [14] but without the log operation. Notice that if the first order transitions are exactly the same between the template and the test segment transition matrix, then CC=1; conversely, if the two have no similarity at all then CC=0, so it is intuitive to name this quantity as 'correlation coefficient'. Note however that CC(Z,Y) is always positive.

3 Simulation

Known time series were created from the Mackey Glass and Lorenz models using the Runge-Kutta method with integration step 0.01, sampling rate 6. First 60,000 samples of Mackey Glass (MG) series with τ=30, initial values 0.9 was generated. The embedding parameters were decided using the method introduced in [15], which yields an embedding dimension of m=6, delay L=2. The first 5,000 samples of this time series were used to train a MG30 SOM template (the SOM is two dimensional with 20 by 20 PEs. this training section is also used as

the reference section of box method and Diks test). The rest of MG30 was concatenated in the middle of other Mackey Glass series with different τ =17, 25, 27, 28, as well as with a Lorenz series x variable with σ=10 (x variable), r=28 (y variable), b=8/3 (z variable), initial conditions x(0)=1, y(0)=0, z(0)=0. All time series are normalized to the range [-0.5 0.5] before template training and mixing. So the testing sets are Lorenz(x)-MG30-Lorenz(x), MG17-MG30-MG17, MG25-MG30-MG25, MG27-MG30-MG27, MG28-MG30-MG28.



Trained SOM template
Voronoi Cell Center

1 Nodes on the trajectory

Density distribution Q_j

Voronoi Cell	(a)	(b)	(c)
Density	5/14	4/14	5/14

'State' transition distribution Y_{ij}

Distribution / From	To	(a)	(b)	(c)
(a)		2/13	2/13	0
(b)		0	2/13	2/13
(c)		2/13	0	3/13

Figure1.The density distribution Q_j and transition distribution Y_{ij} of a trained template

Figure 3 (a) is the Kullback-Leibler distance generated by the proposed method, figure 3 (b) is the Diks' distance between the same test series and the MG30 template time series, figure 3 (c) is the distance generated by box method. Notice that the numerical distance is insignificant for Diks method, as it depends on the window size of test.

The significance of our results is contained in the drop observed when the MG30 is encountered. The SOM template method provides better quantification of the differences than the box method, which is very selective but it is unable to quantify well similarity. Indeed, the box method drops clearly when MG 30 pattern is found, but provides a weak discrimination of the differences in patterns close to MG 30 (i.e. the sensitive of box methods is reduced). Note that in this respect the SOM template method behaves very much like Diks test. However, the SOM based method looses some sensitivity for MG 29 which is also consistent with the unavoidable quantization error. Figure 4 depicts the corresponding first order dynamic similarity of the same test data generated by SOM method. As we can see the results are also very similar for this case and seem not to provide any extra advantage.

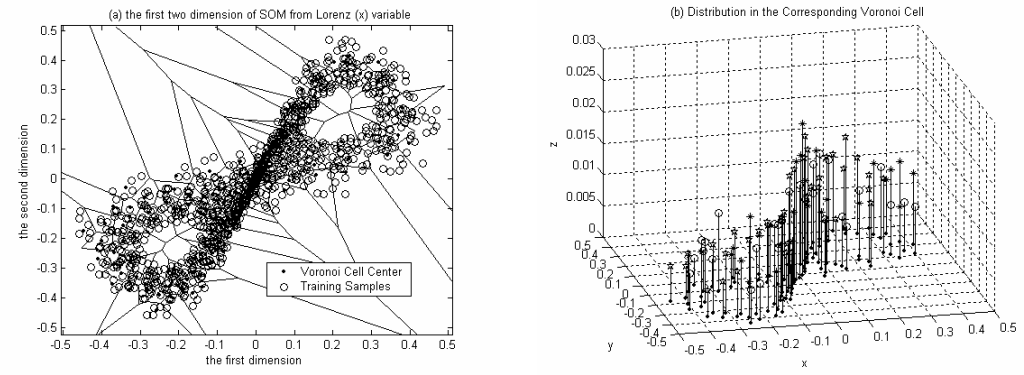


Figure 2 The first two dimension of Lorenz (x) SOM template
(a).Voronoi cell center $\boldsymbol{W_{js}}$ and training samples
(b).Sample distribution $\boldsymbol{Q_{js}}$ in each corresponding Voronoi cell

4 Summary

In this paper we proposed a method to quantify similarity between known time series patterns and time series under testing. The conditions to make the clustering of time series meaningful as had been discussed in [16][17] are taken care of by the embedding process before similarity testing. In stead of using sliding windows, the embedding is based on the assumption that the time series under study was generated by a nonlinear dynamic system. Armed with this assumption the embedding dimension and delay was carefully chosen through established procedures as described in [15] [18].

In essence the proposed method is to measure similarity among trajectories of the time series. But instead of doing this in the original reconstruction space or on the time series, we propose to use a SOM to project and discretize the trajectory to a 2 D space. Two distance measures are proposed and compared: the KL divergence (not limited to KL, any divergence measure based on density distribution will do) which is a static measure of the density of points, and the correlation coefficient measure that is also sensitive to the time evolution of the trajectory.

We compared the performance of this approach with Diks' distance and the simple box counting method in synthetic time series. The conclusion is that the proposed SOM method performs at the same level as the Diks' test. Therefore, when a time series pattern is known, then the SOM based method can be used to find similar time series segments that might be buried in the test data. The price paid for the proposed method is the training of a SOM template, and the gain is a computation cost on the order of the box counting density estimation methods, but with a much better performance. The SOM template method is therefore a clear winner in the compromise between performance and computational cost.

Acknowledgements: This work was partially supported by NIH R01 EB002089.

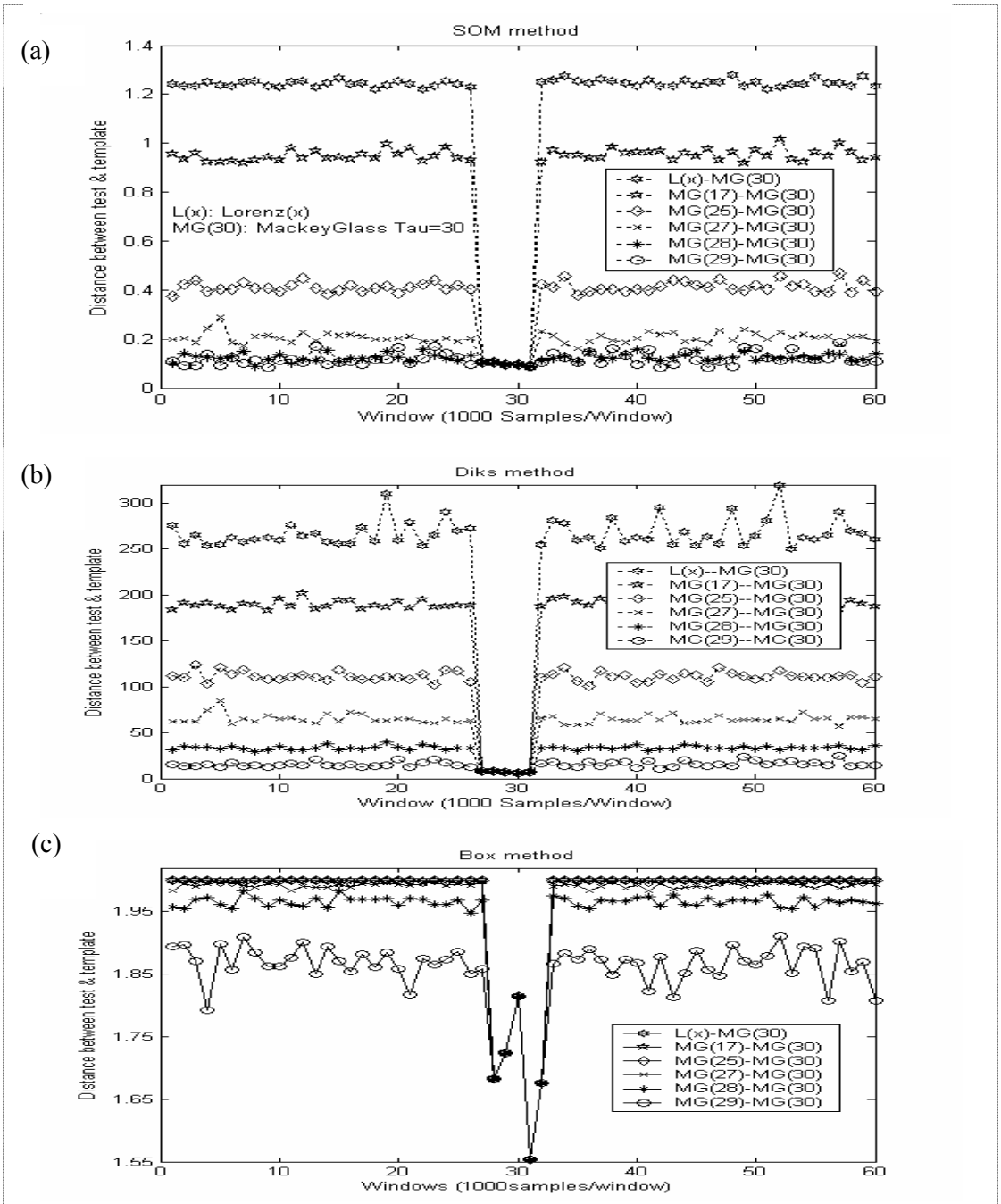


Figure 3 Divergences between MG30 template and the artificial mixed time series
(a). Kullback-Leibler criteria via SOM template method
(b). Diks distance
(c). Distribution distance via box discretization method

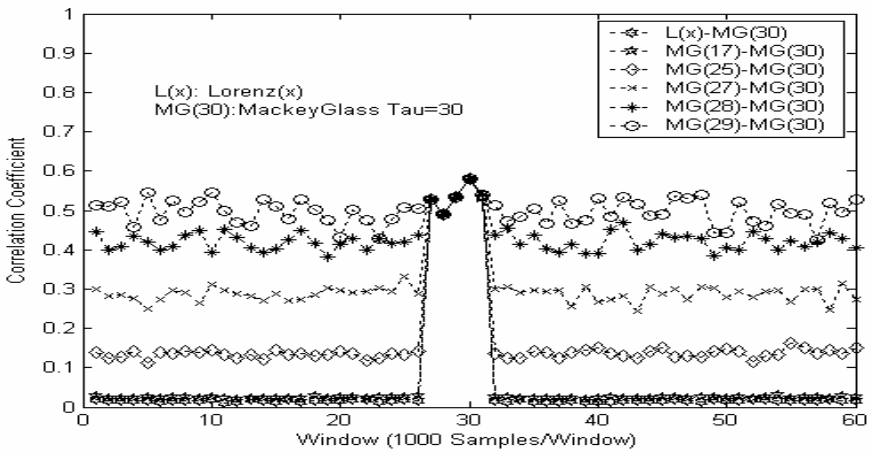


Figure 4.The first order dynamic correlation coefficients between MG30 template and the artificial mixed time series

References

[1].F. Takens (1980), in *Dynamical Systems and Turbulence, Lecture Notes in Mathematics* vol 898, New York, Springer Verlag.

[2].H. Kantz and T. Schreiber (1997), *Nonlinear time series analysis*, London, The Press Syndicate of the University of Cambridge.

[3].P. Grassberger and I. Procaccia (1983), Characterization of strange attractor *Phys. Rev. Lett*. **vol.50**, p.346-349.

[4].A.M. Albano, P.E. Rapp, and A. Passamante(1995), Kolmogorov-Smirnov test distinguishes attractors with similar dimensions. *Phys. Rev. E* ,**vol.52**, 196-206.

[5].T.Schreiber and A. Schmitz (1997), Classification of time series data with nonlinear similarity measures. *Phys. Rev. Lett*. **vol. 79**, p.1475-1478.

[6].H. Kantz (1994), Quantifying the closeness of fractal measures, *Phys. Rev. E* **vol.49**, p.5091-97

[7].A. Wolf, J.B. Swift, H.L. Swinney and J.A. Vastano(1985), Determining Lyapunov exponents from a time series *Physica D* **vol.16**, p.285-317

[8].Yonghong Tan and M. Saif (2000),Neural-networks-based nonlinear dynamic modeling for automotive engines. *Neurocomputing* **vol.30**, 129-142.

[9].J. Wright and R.L. Schult(1993), Recognition and classification of nonlinear chaotic signals *Chaos* **vol.3**, p.295-304.

[10].J. Wright (1995), Monitoring changes in time of chaotic nonlinear systems *Chaos* **vol.5**, p.356-366.

[11].C. Diks, W.R. van Zwet, F. Takens, and J. DeGoede(1996), Detecting difference between delay vector distributions, *Phys. Rev. E* **vol.53**, p.2169-2176 (1996).

[12]. T. Schreiber (1995), Efficient neighbor searching in nonlinear time series analysis, *Int. J. Bifurcation Chaos* **Vol.5** p349

[13].S. Haykin (2000), *Neural Networks: A comprehensive foundation*, Pearson Education Inc.

[14].J.C. Principe, D. Xu, and J.W. Fisher (2000), *Unsupervised adaptive filtering*, John Wiley & Son.

[15].H.D.I. Abarnanel (1996), *Analysis of Observed Chaotic Data*, New York, Springer Verlag.

[16].E. Keogh and Jessica Lin, Clustering of time subsequences is meaningless: implications for previous and future research, *Proc. 3rd Int. IEEE Conf. on Data Mining, Melbourne FL*, p115-122, 2003

[17]. Z. Struzik, Time series rule discovery: tough, not meaningless, *Int. Symp. On Methodologies for Intelligent Systems*, Japan, LNCS Vol.2871, p32-39 Springer 2003

[18]. R. C. Hilborn (2000), *Chaos and Nonlinear Dynamics*, Oxford University Press.