Mining Time Series Data: A Selective Survey

Marcella Corduas

Abstract Time series prediction and control may involve the study of massive data archive and require some kind of data mining techniques. In order to make the comparison of time series meaningful, one important question is to decide what similarity means and what features have to be extracted from a time series. This question leads to the fundamental dichotomy: (a) similarity can be based solely on time series *shape*; (b) similarity can be measured by looking at time series *structure*. This article discusses the main dissimilarity indices proposed in literature for time series data mining.

1 Introduction

Prediction and control are typical objectives of time series analysis and many applications in real-life involve the study of massive data archive and require some kind of data mining techniques. The real challenge is the large amount of data available which makes any traditional "ad hoc" procedure useless.

For this reason, data mining implies a strong role of data processing and the related research field has been significantly occupied by researchers working on database management and machine learning who have often rediscovered known statistical techniques and rarely considered the inferential problem (see Keogh and Kasetty 2003, for a review).

In this article the attention will be focused on the *indexing* problem which, given a time series (a query sequence), finds the nearest matching time series in a database. The solution is achieved in two steps: firstly, a subset of series is selected by means of a crude or approximate dissimilarity criterion; secondly, a refined search is performed. In this respect, the leading idea is that data mining techniques have to discover objects that move similarly or closely follow certain given pattern. This

M. Corduas
Dipartimento di Scienze Statistiche, Università di Napoli Federico II, Via L.Rodino, 80138, Napoli(I), Italy
e-mail: corduas@unina.it

F. Palumbo et al. (eds.), *Data Analysis and Classification*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-03739-9_40, © Springer-Verlag Berlin Heidelberg 2010

concept is typical of shape based dissimilarity measures. However, as this article will discuss, the final objective of a statistical analysis may lead to different approaches where time series modelling assumes a definite role. The article is organized as follows: in Sect. 2, the dissimilarity measures based on shape comparison are introduced; then, in Sect. 3 the problem of features extraction will be examined both in time and frequency domain. Finally, some distance criteria which compare time series dynamics by looking at the underlying generating processes are considered.

2 Comparing Time Series Shape

The most common device used in practice for data mining purposes is the Euclidean distance between the observations:

$$D_E(x_t, y_t) = \{\sum_{t=1}^{n} [x_t - y_t]^2\}^{1/2}, \tag{1}$$

where x_t and y_t, $t = 1, 2, \ldots n$, are zero mean time series. The distance may be referred to standardized data $\widetilde{x}_t$ and $\widetilde{y}_t$ leading to a more meaningful criterion which is invariant for linear transformation of data. In such a case, D_E^2 is just a linear transformation of the correlation coefficient of the two series $r_{xy}(0)$, being $D_E(\widetilde{x}_t, \widetilde{y}_t) = \{2n(1 - r_{xy}(0)\}^{1/2}$.

From a computational point of view, the distance (1) is very simple to implement since a possible matching candidate to a given time series is dismissed as soon as the distance between the first k observations is larger than a fixed threshold. However, the use of data-base archives may be inefficient since only time series with the same length can be considered. Moreover, the criterion is very sensitive to outliers and to distortion in time axis. The latter implies that the similarity of sequences which are locally out of phase is not detectable. For this reason, in spite of the computational complexity, the Dynamic Time Warping (DTW), originally introduced for speech processing (Sakoe and Ciba 1978; Berndt and Clifford 1994) was reconsidered for data mining purposes. In a certain sense, DTW generalizes the concept of dissimilarity between time series trajectories since it allows non-linear alignments of data. Specifically, given two data sequences $x = \{x_i, i = 1, 2, \ldots, m\}$, and $y = \{y_j, j = 1, 2, \ldots, n\}$, the procedure starts by constructing the $m \times n$ matrix Δ where the (i, j) element is the distance (or dissimilarity) $\delta(x_i, y_j)$ between two points x_i and y_j. The best matching is found by searching a path through this matrix such that the total cumulative distance between the aligned elements of the two time series is minimized.

We denote by $w = \{(i(k), j(k)), k = 1, \ldots, K, i(1) = j(1) = 1, i(K) = m, j(K) = n\}$ with $\max(m, n) \leq K \leq m + n - 1$, a warping path connecting $(1, 1)$ and (m, n). The alignment between the time series is obtained by searching for the path through the matrix Δ which minimizes a cost function such as:

$$\mathcal{C}(x, y, w) = \sum_{k=1}^{K} \delta(x_{i(k)}, y_{j(k)}) r(k), \tag{2}$$

where $r(k)$ is an appropriate non negative weighting function (this is often set to $1/k$). Of course, the choice of the cost function determines the warping result.

Some constraints are imposed in order to reduce the number of paths considered:

- *Boundary*: $i(1) = j(1) = 1, i(K) = m, j(K) = n$
- *Monotonicity*: $i(k) \leq i(k+1)$ and $j(k) \leq j(k+1)$
- *Continuity*: $i(k+1) - i(k) \leq 1$ and $j(k+1) - j(k) \leq 1$
- *Window*: the path is allowed to move within a definite region around the matrix diagonal (see Sakoe and Chiba 1978, for the rectangular window; Itakura 1975, for the parallelogram window)
- *Slope*: the path should be neither too steep nor too shallow.

At the end of the optimizing process, the optimal path provides a measure of the *dynamic warping distance* between the two time series:

$$DTW(x, y) = \inf_{w} \mathcal{C}(x, y, w). \tag{3}$$

The main disadvantages of this technique are the computing burden, which limits its usage in practice, and the sensitivity to extreme data. As a matter of facts, DTW can be severely affected by outliers since it tends to adjust extreme data in one of the time series by relating them to extreme values of the other. Various developments of this technique have been proposed such as, among the others, the derivation of a lower bound for DTW using different types of windows (Keogh 2002; Ratanamahatana and Keogh 2004), the extension to multidimensional data (Vlachos et al. 2006b), the use of smoothing for noisy data (Morlini 2005), the joint use of DTW and *S*elf *O*rganizing *M*ap (SOM) algorithm for improving time series clustering (Romano and Scepi 2006), the study of new techniques for approximating DTW (Chu et al. 2002).

3 Criteria Based on Fourier and Wavelet Analysis

Fourier and wavelet analysis provide a good framework in order to extract time series features which become object of the subsequent comparison.

Firstly, Agrawal et al. (1994) considered the *D*iscrete *F*ourier *T*ransform (DFT) of the data:

$$x(\omega_j) = n^{-1/2} \sum_{t=0}^{n-1} x_t \exp(-\imath \omega_j t), \tag{4}$$

where $\omega_j = 2\pi j/n$, $j = 0, 1, \ldots, (n-1)$ and introduced the criterion:

$$D^2_{A,n} = \sum_{j=0}^{n-1} |x(\omega_j) - y(\omega_j)|^2. \tag{5}$$

Of course, $D^2_{A,n} = D^2_E(x_t, y_t)$, but for indexing purposes, only the first k coefficients of each time series DFT are stored so that a selection of potential candidates for the final matching is simply given by: $D_{A,k} < \epsilon$. Standardizing the data first will allow for differences in level and scale. In this respect, the mentioned indexing strategy has two critical issues: the selection of the threshold (which is data dependent) and, above all, the assumption that low frequencies will be in general more informative about the temporal dynamics.

Other criteria in frequency domain have been investigated such as the Euclidean distance between periodograms (Caiado et al. 2006), periodograms of the standardized series at dominant frequencies (Vlachos et al. 2006a) or between smoothed periodograms (Wang and Wang 2000).

The assumption of stationarity of the data generating process is, in our opinion, a critical issue for all methods which rely on time series features such as periodogram, spectrum or autocorrelation functions. In case of non stationary series, these criteria are still useful whenever the non stationarity is removed from each time series by means of the same differencing operator or detrending technique.

In order to find a valid alternative to the traditional Fourier analysis several contributions have explored the use of wavelet analysis (see, for instance, Struzik and Sieber 1999; Li et al. 2002 for an extensive review). Wavelet transforms (or coefficients) are, in fact, characteristic of the local behaviour of a function whereas Fourier transforms relate to the global behaviour.

Any function $f(t) \in L^2(\Re)$ can be written as a wavelet series expansion:

$$f(t) = \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} w_{j,k} \psi_{j,k}(t), \tag{6}$$

where the set of basis functions:

$$\{\psi_{j,k}(t) = 2^{(j/2)} \psi(2^j t - k), \;\; j, k \in Z\} \tag{7}$$

are obtained by dilations and translations of a mother wavelet $\psi(t)$ and the coefficients are given by: $w_{j,k} = \int_{-\infty}^{\infty} f(t) \psi_{j,k}(t) dt$.

Data mining mainly relies on the use of the Haar *D*iscrete *W*avelet *T*ransform (DWT) which simply extracts the underlying pattern of a time series by a recursive pairwise averaging and differencing of data. Chan and Fu (1999) showed the preservation of Euclidean distance in both time and Haar domain and, by analogy to other mining techniques mentioned above, they proposed to retain only the first few coefficients of the transformed sequences in order to perform a similarity search. Later, the approach was improved by taking the "best" (that is, largest) Haar coefficients into account.

4 Structural Dissimilarity

The interest for the dynamic structure inevitably conveys the investigation to the stochastic generating process that has originated the observed trajectory.

In this respect, the class of Gaussian $ARIMA$ processes provides a useful parsimonious representation (Box and Jenkins 1976) for linear time series. Specifically, $Z_t \sim ARIMA(p, d, q)$ is defined by:

$$\phi(B)\nabla^d Z_t = \theta(B)a_t, \tag{8}$$

where a_t is a Gaussian White Noise (WN) process with constant variance $\sigma^2_{a_z}$, B is the backshift operator such that $B^k Z_t = Z_{t-k}$, $\forall k = 0, \pm 1, \ldots$, the polynomials $\phi(B) = 1-\phi_1 B-\ldots-\phi_p B^p$ and $\theta(B) = 1-\theta_1 B-\ldots-\theta_q B^q$, have no common factors, and all the roots of $\phi(B)\theta(B) = 0$ lie outside the unit circle. Moreover, we assume that the time series has been preliminary transformed in order to improve Gaussianity, to deal with non-linearities, to reduce asymmetry, and to remove any outlier or deterministic components (such as deterministic seasonality, trading days, calendar effects, mean level, etc.).

First of all, we will introduce a distance criterion based on the *cepstral coefficients*, $c_{x,j}$, of zero mean stationary series determined by the following expansion:

$$\ln f_x(\omega) = \sum_{j=-\infty}^{\infty} c_{x,j} \exp(-\imath\omega j), \tag{9}$$

where $f_X(\omega)$, $\omega \in (-\pi, \pi]$ is the spectrum of the process X_t (Bogert et al. 1962).

For a pure stationary $AR(p)$ model, a simple expression of the cepstral coefficients in terms of the AR parameters can be derived (Gray and Markel 1976). For this reason, for several decades, the cepstral distance:

$$D_{C,k} = \sqrt{\sum_{j=1}^{k} \left[c_{x,j} - c_{y,j}\right]^2} \tag{10}$$

had been widely applied to signal processing (see for instance Markel and Gray 1976; Kang et al. 1995), and, more recently, it was used for data mining purposes (Kalpakis et al. 2001).

Note that, in the expression (12), the term $(c_{x,0} - c_{y,0})^2 = \ln(\sigma^2_{a_x}/\sigma^2_{a_y})$ is omitted since it is the log of the White Noise variance ratio and hence it simply represents a scale factor. Moreover, the cepstral coefficients quickly decay to zeros, and then, by analogy to previous methods, just a few number of cepstral coefficients, M, have to be stored for indexing purposes so that the Euclidean distance will be computed on the truncated series of cepstral coefficients.

Several improvements have been proposed for speech recognition purposes such as the use of Mahalanobis distance or the introduction of a weighted Euclidean

distance in which each coefficient is simply weighted by the inverse of its variance in order to enhance the contribution of weights with lower variability (Tohkura 1987).

Furthermore, Piccolo (1984, 1990) proposed a distance criterion which compares the forecasting functions of two $ARIMA$ models given a set of initial values. In particular, assuming that Z_t is a zero mean invertible process which admits the $AR(\infty)$ representations: $\pi(B)Z_t = a_t$, the π-weights sequence and the WN variance completely characterize Z_t (given the initial values). Hence, a measure of structural diversity between two $ARIMA$ processes with given orders, X_t and Y_t, can be defined as:

$$D_{AR} = \sqrt{\sum_{j=1}^{\infty} \left(\pi_{xj} - \pi_{yj}\right)^2}. \tag{11}$$

As before, the WN variances are not included in the distance formulation since they depend on the units of measurement. The criterion has been widely experimented (see Piccolo 2007 for a review) and the asymptotic properties has been derived under general assumptions (Corduas 2000; Corduas and Piccolo 2008). Moreover, Baragona and Vitrano (2007) compared the performance of the AR metric with a criterion based on cross-correlations for data mining purposes.

Recently, Bagnall and Janaceck (2005) suggested to translate a time series into binary sequences in order to reduce the amount of storage and computational resources for time series comparison, and then, to apply the AR metric for subsequent clustering. The technique achieved a clustering accuracy equivalent to that obtained by cepstral distance and proved to be of help in presence of outliers.

In the same vein, the Mahalanobis distance between AR processes was proposed and, as we will discuss in the next section, the related distributional properties were investigated (see Thomson and De Souza 1985, and references therein reported). Xiong and Yeung (2004), instead, introduced a model-based clustering approach based on mixtures of $ARMA$ models.

5 Final Remarks

Concluding this brief review, we illustrate some results which helps the set up of time series comparison within an inferential framework. Assuming that X_t and Y_t are independent Gaussian and stationary zero mean processes, the following results hold.

- *Mahalanobis distance between $AR(p)$ processes*
  The null hypothesis is $H_0 : \boldsymbol{\phi}_x = \boldsymbol{\phi}_y = \boldsymbol{\phi}$, $\sigma^2_{a_x} = \sigma^2_{a_y} = \sigma^2$. Under H_0, $(\hat{\boldsymbol{\phi}}_x - \hat{\boldsymbol{\phi}}_y) \sim N_p(0, 2n^{-1}\sigma^2\boldsymbol{\Gamma}^{-1})$, being $\hat{\boldsymbol{\phi}}_x$ and $\hat{\boldsymbol{\phi}}_x$ the ML estimator vector of AR parameters, and Γ the p-order Toeplitz matrix of the common generating process. Hence,

$$\widehat{M}^2(X_t, Y_t) = \frac{n}{2}(\hat{\boldsymbol{\phi}}_x - \hat{\boldsymbol{\phi}}_y)'\sigma^{-2}\boldsymbol{\Gamma}(\hat{\boldsymbol{\phi}}_x - \hat{\boldsymbol{\phi}}_y)$$

is asymptotically distributed a $\chi_{(p)}$ random variable. When the matrix $\sigma^{-2}\mathbf{\Gamma}$ is unknown, it will be replaced by the corresponding pooled estimator. Also, the criterion can be generalized to time series with different length (Thomson e De Souza 1985).

- *AR metric*
  The null hypothesis $H_0 : D^2_{AR} = 0$, is equivalent to: $H_0 : \boldsymbol{\pi}_x - \boldsymbol{\pi}_y = \mathbf{0}$. For $ARMA$ processes, the m-lag truncated ML estimator $\widehat{D}^2_m$ is asymptotically distributed as a linear combination of independent χ^2_1 random variables; the weights are the non zero eigenvalues of $\mathbf{C_0} = \left(\frac{1}{n_x} + \frac{1}{n_y}\right)\mathbf{BVB}'$, being $\mathbf{V}$ the covariance matrix of the ML estimators of $ARMA$ parameters and $\mathbf{B} = \{b_{ij}\}$ with $b_{ij} = \left\{\partial\widehat{\pi}_i/\partial\widehat{\beta}_j\right\}_{\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}}$, where $\beta_j = \phi_j$, $j = 1,\ldots,p$, and $\beta_{j+p} = \theta_j$, $j = 1,\ldots,q$. The result can be easily generalized to $ARIMA$ processes (Corduas 2000; Corduas and Piccolo 2008).

Other dissimilarity criteria have been proposed in the statistical literature which have not found a clear role so far, although the fields of application, which they were originally designed for, typically require the use of large data archives. For instance, we refer to the Kullback-divergence (Shumway 1982), the Bhattacharrya distance (Kazakos and Papantoni-Kazakos 1980) which were largely applied for signal recognition. Moreover, any distance criterion has to be related to an adequate clustering technique in order to produce sensible and useful results. Then, the study of efficient methods for clustering is a further and crucial research topic for data mining (Scepi and Milone 2007).

Acknowledgements This research was supported by Dipartimento di Scienze Statistiche, University of Naples Federico II, and CFEPSR (Portici).

References

Agrawal, R., Faloutsos, C., & Swami, A. (1994). Efficient similarity search in sequence databases. 4th F.O.D.O. *Lecture notes in Computer Science* (Vol. 730, pp. 69–84). New York: Springer

Bagnall, A. J., & Janaceck, G. J. (2005). Clustering time series from ARMA models with clipped data. *Machine Learning*, *58*, 151–178

Baragona, R., & Vitrano, S. (2005). Statistical and numerical algorithms for time series classification. *Proceedings of CLADAG 2007* (pp. 65–68). EUM, Macerata

Berndt, D., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. Proceedings of the AAAI-94 workshop of SIGKDD, pp. 229–248

Box, G. E. P., & Jenkins, G. M. (1994). *Time series analysis: Forecasting and control* (rev. ed.). San Francisco: Holden-Day

Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, *50*, 2668–2684

Chan, K., & Fu, A. W. (1999). *Efficient time series matching by wavelets*, ICDE (pp. 126–133)

Chu, S., Keogh, E., Hart, D., & Pazzani, M. (2002). Iterative deepining dynamic time warping for time series. Proceedings of SIAM KDD, electronic edition

Corduas, M. (2000). La metrica Autoregressiva tra modelli ARIMA: una procedura in linguaggio GAUSS. *Quaderni di Statistica*, *2*, 1–37

Corduas, M., & Piccolo, D. (2008). Time series clustering and classification by the Autoregressive Metric. *Computational Statistics & Data Analysis*, *52*, 1860–1872

Gray, A. H., & Markel, J. D. (1976). Distance measures for speech recognition. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *ASSP-24*, 380–391

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *ASSP-23*, 67–72

Kalpakis, K., Gada, D., & Puttagunta, V. (2001). Distance measures for effective clustering of ARIMA time series. *IEEE International Conference on Data Mining*, 273–280

Kang, W., Shiu, J., Cheng, C., Lai, J., Tsao, H., & Kuo, T. (1995). The application of cepstral coefficients and maximum likelihood method in EGM pattern recognition. *IEEE Transactions on Biomedical Engineering*, *42*, 777–785

Kazakos, D., & Papantoni-Kazakos, P. (1980). Spectral distances between Gaussian processes. *IEEE Transactions on Automatic Control*, *AC-25*, 950–959

Keogh, E. (2003). Exact indexing of dynamic time warping. *28th International Conference on VLDB* (pp. 406–417). Hong Kong

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*, 349–371

Li, T., Li, Q., Zhu, S., & Ogihara, M. (1997). A survey on wavelet applications in data mining. *SIGKDD Explorations*, *4*, 49–68

Markel, J. D., & Gray, A. H. (1976). *Linear prediction of speech*. New York: Springer

Morlini, I. (2005). On the dynamic time warping for computing the dissimilarity between curves. In M. Vichi, P. Monari, S. Mignani, & A. Montanari (Eds.), *New developments in classification and data analysis* (pp. 63–70). Berlin: Springer

Piccolo, D. (1984). Una topologia per la classe dei processi ARIMA. *Statistica*, *XLIV*, 47–59

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, *11*, 153–164

Piccolo, D. (2007). Statistical issues on the AR metric in time series analysis. *Proceedings of the SIS Intermediate Conference* (pp. 221–232). Cleup, Padova

Ratanamahatana, C. A., & Keogh, E. (2004). Making time-series classification more accurate using learned constraints. *4-th SIAM International Conference on Data Mining* (pp. 1–20)

Romano, E., & Scepi, G. (2004). Integrating time alignment and Self-Organizing Maps for classifying curves. Proceedings of KNEMO COMPSTAT 2006 Satellite Workshop, Capri

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *26*, 143–165 (1978).

Scepi, G., & Milone, G. (2007). Temporal data mining: clustering methods and algorithms. *Proceedings of CLADAG 2007* (pp.73–76). EUM, Macerata

Shumway, R. H. (1982). Discriminant analysis for time series. In Krishnaiah,P. R. & Kanals, L.N. (Eds.), *Handbook of Statistics* (Vol. 2, pp. 1–46). New York: North Holland

Struzik, Z. R., & Siebes, A. (1999). The Haar wavelet in the time series similarity paradigm. *3rd European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 12–22). Prague: Springer

Thomson, P. J., & De Souza, P. (1985). Speech recognition using LPC distance measures. In E. J. Hannan, P. R. Krishnaiah, & M. M. Rao (eds.), *Handbook of Statistics* (Vol. 5, pp. 389–412). Amsterdam: Elsevier

Tohkura, Y. (1987). A weighted cepstral distance measure for speech recognition. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, *ASSP-35*, 1414–1422

Vlachos, M., Yu, P., Castelli, V., & Meek, C. (2006a). Structural periodic measures for time series data. *Data Mining and Knowledge Discovery*, *12*, 1–28

Vlachos, M., Hadejieleftheriou, M., Gunopulos, D., & Keogh, E. (2006b). Indexing multidimensional time series. *The VBDL Journal*, *15*, 1–20

Wang, C., & Wang, X. S. (2000). Supporting content-based searches on time series via approximations. 12th International Conference on Scientific and Statistical Database Management (pp. 69–81)

Xiong Y., & Yeung D. (2004). Time series clustering with ARMA mixtures. *Pattern Recognition*, *37*, 1675–1689