Improved Fréchet Distance for Time Series

Ahlame Chouakria Douzal, P. Nagabhushan

Improved Fréchet Distance For Time Series

A. Chouakria-Douzal[1] and P. Nagabhushan[2]

[1] TIMC-IMAG, Université Joseph Fourier Grenoble 1,
F-38706 LA TRONCHE Cedex, France
Ahlame.Douzal@imag.fr
[2] Dept. of Studies in Computer Science, University of Mysore
Manasagangothri, Mysore, Karnataka- 570 006, India
pn@amrita.edu

Abstract. This paper focuses on the Fréchet distance introduced by Maurice Fréchet in 1906 to account for the proximity between curves (Fréchet (1906)). The major limitation of this proximity measure is that it is based on the closeness of the values independently of the local trends. To alleviate this set back, we propose a dissimilarity index extending the above estimates to include the information of dependency between local trends. A synthetic dataset is generated to reproduce and show the limited conditions for the Fréchet distance. The proposed dissimilarity index is then compared with the Fréchet estimate and results illustrating its efficiency are reported.

1 Introduction

Time series differ from "non-temporal" data due to the interdependence between measurements. This work focuses on the distances between time series, an important concept for time series clustering and pattern recognition tasks. The Fréchet distance is one of the most widely used proximity measure between time series. Fréchet distance uses time distortion by acceleration or deceleration transformations to look for a mapping that minimizes the distance between two time series. We show in section 4, that the Fréchet distance ignores the interdependence among the occurring values; proximity is only based on the closeness of the values; which can lead to irrelevant results. For this reason, we propose a dissimilarity index extending this classical distance to include the information of dependency between local trends. The rest of this paper is organized as follows: the next section presents the definitions and properties of the conventional Fréchet distance. Section 3, discusses the major limitations of such proximity estimate, then gives the definition and properties of the new dissimilarity index. Section 4, presents a synthetic dataset reproducing limited conditions for this widely used time series proximity measure, then perform a comparison between the proposed dissimilarity index and the Fréchet distance before concluding.

2 The Fréchet distance between Time Series

The success of a distance, intended to distinguish the events of a time series that are similar from those that are different, depends on its adequacy with respect to the proximity concept underlying the application domain or the experimental context.

The Fréchet distance was introduced by Maurice Fréchet in 1906 (Fréchet (1906)) to estimate the proximity between continuous curves. We present a discrete variant of this distance. An in-depth study of the Fréchet distance is provided by Alt (Alt and Godau (1992)) and an interesting comparison of the different distance theories can be found in Eiter and Mannila (1994). The popular and highly intuitive Fréchet distance definition is: "A man is walking a dog on a leash. The man can move on one curve, the dog on another. Both may vary their speed independently, but are not allowed to go backwards. The Fréchet distance corresponds to the shortest leash that is necessary". Let's provide a more formal definition.

We define a mapping $r \in M$ between two time series $S_1 = (u_1, ..., u_p)$ and $S_2 = (v_1, ..., v_q)$ as the sequence of m pairs preserving the observation order:

$$r = ((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m}))$$

with $a_i \in \{1, .., p\}$, $b_j \in \{1, .., q\}$ and satisfying for $i \in \{1, .., m-1\}$ the following constraints:

$$a_1 = 1,\ a_m = p \qquad b_1 = 1, b_m = q \tag{1}$$

$$a_{i+1} = a_i \ \text{ or } \ a_i + 1 \qquad b_{i+1} = b_i \ \text{ or } \ b_i + 1 \tag{2}$$

We note $|r| = \max_{i=1,..,m} |u_{a_i} - v_{b_i}|$ the mapping length representing the maximum span between two coupled observations. The Fréchet distance $\delta_F(S_1, S_2)$ is then defined as:

$$\delta_F(S_1, S_2) = \min_{r \in M} |r| = \min_{r \in M} (\max_{i=1,..,m} |u_{a_i} - v_{b_i}|) \tag{3}$$

Graphically, a mapping between two time series $S_1 = (u_1, ..., u_p)$ and $S_2 = (v_1, ..., v_q)$ can be represented by a path starting from the corner $(1, 1)$ and reaching the corner (p, q) of a grid of dimension (p, q). The value of the square (i, j) is the span between the coupled observations (u_i, v_j). The path length corresponds to the maximum span reached through the path. Then, the Fréchet distance between S_1 and S_2 is the minimum length through all the possible grid paths. We can easily check that δ_F is a metric verifying the identity, symmetry and triangular inequality properties (a proof can be found in Eiter and Mannila (1994)).

According to δ_F two time series are similar if there exists a mapping between their observations, expressing an acceleration or a deceleration of the occurring observation times so that the maximum span between all coupled observations is close.

Note that the Fréchet distance is very useful when only the occurring events, not their occurring times, are determinant for the proximity evaluation. This explains the great success of Fréchet distance in the particular domain of voice processing where only the occurring syllables are used to identify words; the flow rate being specific to each person.

3 Fréchet distance Extension for Time Series Proximity Estimation

Generally, the interdependence among the occurring values, characterizing the local trends in the time series, is determinant for the time series proximity estimation. Thus, Fréchet distance fails as it ignores such main information. Section 4 illustrates two major constraints in the Fréchet measure: ignorance of the temporal structure and the sensitivity to global trends. To alleviate these drawbacks in the classical Fréchet estimate we propose a dissimilarity index extending Fréchet distance to include the information of dependency between the time series local trends. The dissimilarity index consists of two components. The first one estimates the closeness of values and is based on a normalized form of the conventional proximity measure. The second component, based on the temporal correlation Von Neumann (1941-1942)), Geary (1954) and (Chouakria-Douzal (2003), estimates the dependency between the local trends.

3.1 Temporal Correlation

Let's first recall the definition of the temporal correlation between two time series $S_1 = (u_1, ..., u_p)$ and $S_2 = (v_1, ..., v_p)$:

$$\text{CORT}(S_1, S_2) = \frac{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)^2 \sum_{i=1}^{p-1}(v_{(i+1)} - v_i)^2}}$$

The temporal correlation coefficient $\text{CORT} \in [-1, 1]$ estimates how much the local trends observed simultaneously on both times series, are positively/negatively dependent. By dependence between time series we mean a stochastic linear dependence: if we know at a given time t the growth of the first time series then we can predict, through a linear relationship, the growth of the second time series at that time t. Similar to the classical correlation coefficient, a value of $\text{CORT} = 1$ means that, at a given time t, the trends observed on both time series are similar in direction and rate of growth, a value of -1 means that, at a given time t, the trends observed on both time series are similar in rate but opposite in direction and finally, a value of 0 expresses that the trends observed on both time series are stochastically linearly independent.

Contrary to the classical correlation coefficient, the temporal correlation estimates locally not globally the dependency between trends; indeed, two time series may be highly dependent through the classical correlation and linearly independent through the temporal correlation (illustrated in section 4). Finally, contrary to classical correlation, the temporal correlation is global trend effect free. Let's now present the new dissimilarity index as an extension of the Fréchet distance.

3.2 The dissimilarity Index

The proposed dissimilarity index consists in the combination of two components. The first one, estimates the closeness of values and is based on a normalized form of the Fréchet distance. The second one is based on the temporal correlation introduced above. Many functions could be explored for such combination function. To illustrate well the additive value of the temporal correlation to account for local trends dependency, we limit this work to a linear combination function. Let's note $DisF$ the dissimilarity index extending δ_F:

$$DisF(S_1, S_2) = \alpha \left(\frac{\delta_F(S_1, S_2)}{\max_{S_i, S_j \in \Omega_S} \delta_F(S_i, S_j)} \right) + (1 - \alpha) \left(\frac{1 - \text{CORT}(S_1, S_2)}{2} \right)$$

where $DisF(S_1, S_2) \in [0, 1]$, Ω_S is the set of the observed time series, $\alpha \in [0, 1]$ determines the weight of each component in the dissimilarity evaluation and CORT the temporal correlation defined above.

Note that for $\alpha = 1$, $DisF$ corresponds to the normalized δ_F and the proximity between two time series is only based on taken values, considered as independent observations. For $\alpha = 0$, $DisF$ corresponds to CORT and the proximity is based solely on the dependency between local trends. Finally for $0 < \alpha < 1$, $DisF$ implies a weighted mean of the normalized δ_F and CORT, the proximity between time series includes then, according to their weights, both the proximity between occurring values and the dependency between local trends.

4 Applications and Results

In this section, we first present the time series synthetic dataset which reproduces the limited conditions for Fréchet distance. Then we explore and compare the distribution of the temporal and classical correlations between the synthetic dataset time series. Finally, the proposed dissimilarity index is compared to the conventional estimate.

4.1 Synthetic Dataset

To reproduce the limited conditions for the widely used conventional distances, we consider a synthetic dataset of 15 time series divided into three

classes of functions. The first five time series are of class F_1, the next five are of class F_2 and the last five are of the class F_3; where, F_1, F_2 and F_3 are defined as follows:

$$F_1 = \{f_1(t) \mid f_1(t) = f(t) + 2t + 3 + \epsilon\}$$
$$F_2 = \{f_2(t) \mid f_2(t) = \mu - f(t) + 2t + 3 + \epsilon\}$$
$$F_3 = \{f_3(t) \mid f_3(t) = 4f(t) - 3 + \epsilon\}$$

$f(t)$ is a given discrete function, $\mu = E(f(t))$ is the mean of $f(t)$ through the observation period, $\epsilon \rightsquigarrow N(0, 1)$ is a zero mean gaussian distribution and 2t + 3 describes a linear upward trend tainting F_1 and F_2 classes. Figure 1 represents simultaneously these three classes through 15 synthetic time series. Note that F_1 and F_3 show similar local tendencies, they increase (respectively

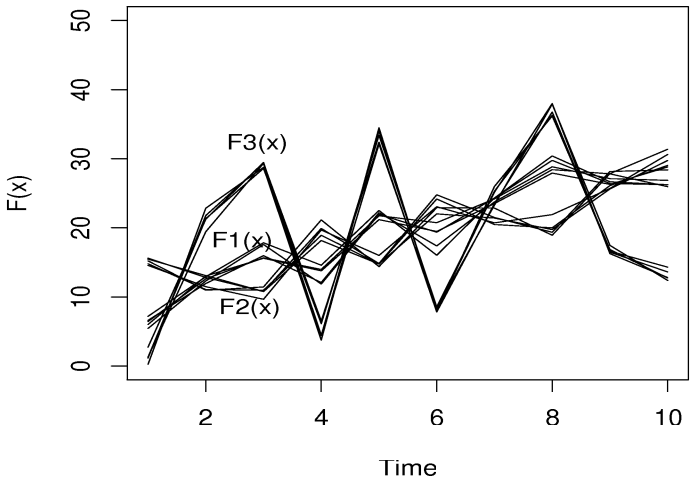

Fig. 1. Three classes of synthetic time series

decrease) simultaneously. On the contrary, F_2 shows local tendencies opposite to those of F_1 and F_3, when F_2 increases (respectively decreases) F_1 and F_3 decreases (respectively increases). Finally, F_1 and F_2 are the closest in values.

4.2 Time Series Temporal correlation vs Classical correlation

Let's explore in figure 2 the distribution of the temporal and classical correlations among the times series into F_1, F_2 and F_3 classes. On the one hand, the

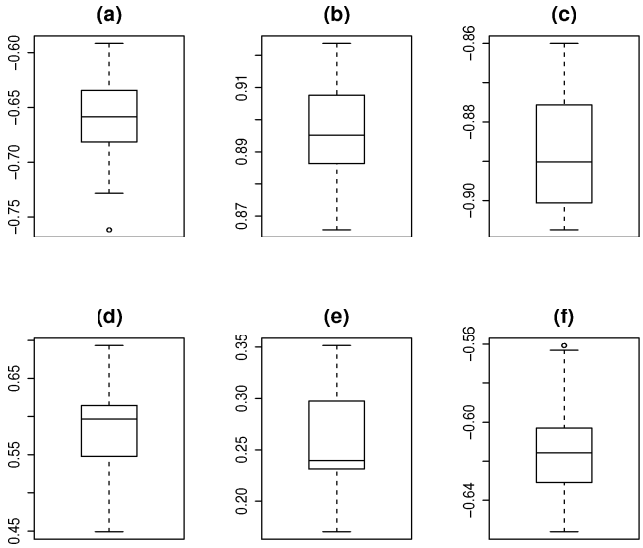

Fig. 2. **(a)** $\text{CORT}(F_1(x), F_2(x))$ **(b)** $\text{CORT}(F_1(x), F_3(x))$ **(c)** $\text{CORT}(F_2(x), F_3(x))$ **(d)**$\text{COR}(F_1(x), F_2(x))$ **(e)** $\text{COR}(F_1(x), F_3(x))$ **(f)** $\text{COR}(F_2(x), F_3(x))$

temporal correlation distribution $\text{CORT}(F_1, F_3) \in [0.87, 0.92]$, $\text{CORT}(F_1, F_2) \in [-0.73, -0.60]$ and $\text{CORT}(F_2, F_3) \in [-0.91, -0.86]$) reveal a high positive dependency between F_1 and F_3 classes and a high negative dependency between F_2 and the two remaining classes. These results supported well the dependencies illustrated above in figure 1.

On the other hand, the classical correlation distribution $\text{COR}(F_1, F_3) \in [0.15, 0.35]$, $\text{COR}(F_1, F_2) \in [0.45, 0.70]$ and $\text{COR}(F_2, F_3) \in [-0.66, -0.56]$) indicates a weak (nearly independence) positive dependency between F_1 and F_3 classes and a high positive dependency between F_1 and F_2 classes. These results illustrate well that the classical correlation estimates globally (not locally) the dependency between tendencies of time series. Indeed, F_1 and F_2 which are not locally but globally dependent, due to the linear upward trend tainting them, are considered as highly dependent; whereas F_1 and F_3 which are dependent locally not globally are considered as very weakly dependent. Note that contrary to classical correlation, the temporal correlation is global-trend effect free.

4.3 Comparative Analysis

To compare the above proximity measures, we estimate first the proximity matrices between the 15 synthetic time series, according to $DisF$ and δ_F. $DisF$ is evaluated with $\alpha = 0.5$ and $\alpha = 0$. For $\alpha = 1$, results are

similar to those obtained from δ_F. A hierarchical cluster analysis is then performed on the obtained proximity matrices. Figure 3 illustrates the obtained dendograms. Note first that the three above proximity measures (δ_F,

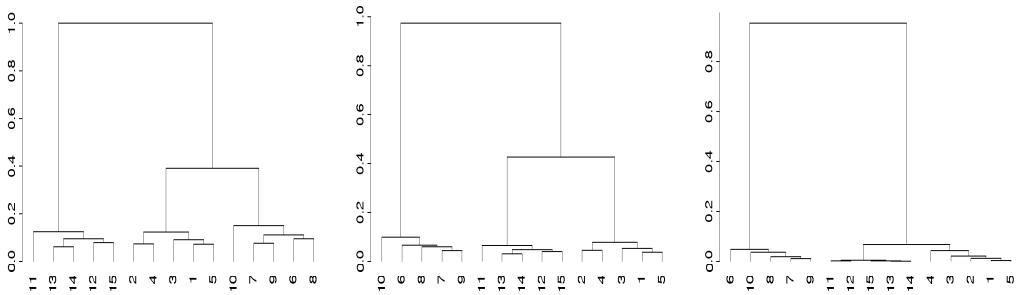

Fig. 3. δ_F $DisF$ $(\alpha = 0.5)$ $DisF$ $(\alpha = 0)$

$DisF(\alpha = 0.5)$ and $DisF(\alpha = 0)$) divide the 15 time series on the well expected three classes F_1 (from 1 to 5), F_2 (from 6 to 10) and F_3 (from 11 to 15). In addition, on the one hand, δ_F dendogram works out the time series of the classes F_1 and F_2 as the closest. Indeed, for δ_F, after stretching each class to match well an other class, the proximity evaluation is based solely on the taken values, which are close on F_1 and F_2.

On the other hand, $DisF$ for $\alpha = 0.5$ and $\alpha = 0$ determines successfully the classes F_1 and F_3 as the closest. Note particularly that for $\alpha = 0.5$ $DisF$ still provides three classes with a high proximity between F_1 and F_3; whereas for $\alpha = 0$ F_1 and F_3 are nearly merged and the respective dendogram comes out with only two main classes. Indeed, for $\alpha = 0$ the proximity evaluation are based solely on the dependency between time series which is very high between F_1 and F_3.

5 Discussion and Conclusion

This paper focuses on the Fréchet distance between time series. We have provided the definitions and properties of this conventional measure. Then we illustrated the limits of this distance. To alleviate these limits, we propose a new dissimilarity index based on the temporal correlation to include the information of dependency between the local trends.

Note that, as this paper introduces the benefits of the temporal correlation for time series proximity estimation, and mainly for clarity reasons, we

limit our work on two points. First we restrict the combination function to a linear function to show clearly, by varying the parameter α, the additive value of the temporal correlation. Secondly, we restrict the illustration of the proposed index to a synthetic dataset which reproduces the limited conditions for the conventional Fréchet distance.

Future works, on the one hand, will study other combination functions. For instance, if we consider the two dimensional space defined by the components CORT and a normalized form of δ_F, then we can define a new euclidean distance between time series as their norm vector in such two dimensional space. On the second hand, these combination functions will be compared to the conventional Fréchet distance through a wide range of a real datasets.

Finally, let's remark that the proposed dissimilarity index $DisF$ could be very useful for time series classification problem, where the aim consists in determining the most adaptable $DisF$ by looking for the optimal value of α maximizing a classification rate. This is an interesting direction to study through a priori time series classification.

References

Geary, R.C. (1954): The contiguity ratio and statistical mapping. *The Incorporated Statistician, 5/3, 115-145.*

Von Neumann, J. (1941): Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics, 12/4.*

Von Neumann, J., Kent,R.H., Bellinson, H.R. and Hart, B.I. (1942): The mean square successive difference to the variance. *The Annals of Mathematical Statistics. 153-162.*

Fréchet, M. (1906): Sur quelques points du calcul fonctionnel. Rendiconti del Circolo Mathematico di Palermo, 22, 1-74.

Godau, M. (1991): A natural metric for curves - computing the distance for polygonal chains and approximation algorithms. In Proc. 8th Sympos. Theor. Aspects of Comp. STACS, LNCS 480, 127-136.

Alt, H. and Godau, M. (1992): Measuring the resemblance of polygonal curves. In Proc. 8th Annu. ACM Sympos. Comput. Geom. 102-109.

Eiter T. and Mannila, H. (1994): Computing Discrete Fréchet distance, Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems. TU Vienna, Austria.

Chouakria-Douzal, A. (2003): Compression Technique Preserving Correlations of a Multivariate Temporal Sequence. In: M.R. Berthold, H-J Lenz, E. Bradley, R. Kruse, C. Borgelt (eds.). *Advances in Intelligent Data Analysis, 5, Springer, 566-577.*