Classification of Multiple Short Time Series

Agostino Tarsitano[1]Dipartimento di Economia Statistica, e-mail: agotar@unical.it

Abstract: The classification of multiple time series (MTS) in groups is an important issue that intersects many research fields. Although much research has been performed in modelling this type data, one area that has been largely overlooked is the particular type of time series where the data set consists of a large number of variables but with a small number of observations. The specific issue addressed in the present project concerns the possibility of classifying a set of administrative units (regions) in clusters on the basis of a new metric defined combining partial metrics computed on selected social and economic indicators that are only available for a limited period of time.

Keywords: Pattern Recognition, Distatis, PAM, Regional analysis.

1. Introduction

Regional analysis often involves multiple time series data because regions may rise and fall at different times and for dissimilar reasons; some sequence has sharp crests or flat troughs, others are similar but there are lags or shifts between the temporal profiles. In fact, modelling several related time series together makes it possible to reveal hidden sequence features, discover development patterns, and, in some situations, develop control schemes. In the context of regional analysis, time series are often of short duration (< 25 time points) because data acquisition has proceeded for but a brief term due to collection problems or high costs.
A classic problem involves clustering the regions into similar categories by extracting distinctive features from the data, compare them and perform the grouping of the regions into distinct categories by means of an appropriate metric. The clustering is judged satisfactory if the distance between objects within clusters is relatively small compared with distances between clusters. Time series clustering is primarily based on the identification of a measure of distance between time series that takes into account the value and ordering of the points along the sequences; then the matrix of interdistances can be used as a raw input into well-known clustering procedures such as hierarchical or partitioning around medoids or PAM methods (**?**. If a region is described by many time-dependent variables then we can obtain as many distance matrices as the number of variables. As a consequence, the comparison of two regions can be considered a comparison of two sets of matrices. In this sense, the Distatis procedure (see **?**) can be very helpful to investigate the relationship between the three-way data generated by the combination of indicators, regions and periods. The paper proceeds as follows. The second section introduces a distance measure for univariate short time series. The third section describes in detail our proposed methods to calculate a distance between multiple time series. Cluster analysis results of regional data are shown in Section 4. Finally, the last section provides a summary and possible lines of future research.

[1]Address of correspondence: Dipartimento di Economia e Statistica, Università della Calabria, Via Pietro Bucci, cubo 1C, Rende (CS) - Italy, Tel.: +39-0984-492465, Fax:+39-0984-492421

2. Distance between univariate time series

Many knowledge discovery and data mining tasks aim to classify a set of regions using time-varying data measured on many process variables over a short period of time. We believe that a method for analyzing such multiple time series should be based on the dissimilarities calculated at the univariate level. There are many ways to define a distance among a set of p univariate time series $\left\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p\right\}$. In some cases, the sequences for territorial units are fairly short and it becomes necessary to exploit the full potential of the time series to assure good generalization ability and better classification accuracy. In this case, one wants more assigning a value to the distance between observed sequences rather than quantifying the strength of relationships between the stochastic processes that generate observations. Consequently, techniques such as autoregressive and spectral representations, decomposition or subsequence matching methods are not useful in case of short length sequences. In a recent review **?**, merits and drawbacks of nine of these measures are discussed and it is pointed out, in particular, that an "appropriate" distance measure should be driven by the relative change of intensity as well as the temporal information. To incorporate both factors, we defined a metric as a weighted average of the separations between intervals

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = d_{i,j} = \frac{\sum_{t=1}^{n} \alpha_t A_t}{\sum_{t=1}^{n} \alpha_t}. \tag{1}$$

where A_t is the area between the two polygons in the time interval $[t, t+1]$. More specifically, A_t is made of trapezoids or triangles. The latter are formed when polygons intersect and the intersection point belongs to $[t, t+1]$. In this case, the local contribution to the distance is obtained by summing the area of the two triangles. In addition, we define

$$\alpha_t = 0.5\left(\min\left\{|x_{i,t} - x_{j,t}|, |x_{i,t+1} - x_{j,t+1}|\right\} + \max\left\{|x_{i,t} - x_{j,t}|, |x_{i,t+1} - x_{j,t+1}|\right\}\right). \tag{2}$$

for $t = 1, 2, \cdots, n-1$. Each α_t represents a time-varying weight assigned to what happens in the interval $[t, t+1]$; the larger is the weight, the more important is the role of the corresponding local distance within the global distance function, while irrelevant differences have virtually zero weight. Of course, the weights in (1) are normalized to the range of zero to one so that $d_{i,j}$ spans from zero to max $A_t, t = 1, 2, \cdots, n$. Practically, the smaller the value of $d_{i,j}$, the greater the similarity between the time series under comparison. The metric (1) has obtained good results in terms recovery of the true group structure.

3. Distance between multivariate time series

Let $\mathbf{X}_r$ be a data matrix $(p \times n_r)$ of observations for a variable measured at n_r time intervals on p units; suppose that similar data exist for m variables. An application of the metric (1) that invokes each of the variables provides a separate and potentially different basis $\mathbf{D}_r, r = 1, 2, \cdots, m$ for clustering the units. We define the r-th partial matrix as

$$\mathbf{D}_r = d_{i,j}^{(r)} \beta_r, r = 1, 2, \cdots, m \quad \text{with} \quad \beta_r = \frac{\max_{r=1,2,\cdots,m}\left\{\max_{i,j=1,2,\cdots p} d_{i,j}^{(r)}\right\}.}{\max_{i,j=1,2,\cdots p} d_{i,j}^{(r)}} \tag{3}$$

The $\beta_r, r = 1, 2, \cdots, m$ are scalars so that the maximum values of $\mathbf{D}_r, r = 1, 2, \cdots, m$ become equal. The scaling procedure is necessary, otherwise the combined representation

would copy the structure of the indicator with the largest distances. To realize a clustering of the MTS we calculated the distances as a weighted average of the univariate distances

$$\mathbf{D} = \sum_{r=1}^{m} \gamma_r \mathbf{D}_r \quad \text{with} \quad \gamma_r \geq 0, \quad \text{and} \quad \sum_{r=1}^{m} \gamma_r = 1. \tag{4}$$

The partial matrices $\mathbf{D}_r, r = 1, 2, \cdots, m$ are, by construction, symmetric and non negative; moreover, since (1) is a metric, the partial matrices are also Euclidean in the sense of **?**. An algorithm that implements (4) can be derived by choosing the weights properly. In this sense, one needs to have an expression for how much a variable affects the global clustering that can be obtained, for example, from the total sum of squares of the elements of the partial distance matrices. This naturally leads to the Distatis procedure. Let $\mathbf{u}_p$ be a $p \times 1$ vector of 1's and $\mathbf{I}$ the identity matrix of order p. A normalized cross-product matrix for $\mathbf{D}_r$ is $\mathbf{T}_r = \theta_{1,r}^{-1}\mathbf{S}_r$ where $\theta_{1,r} > 0$ is the largest eigenvalue of $\mathbf{S}_r = \mathbf{C}\mathbf{D}_r\mathbf{C}^t$ and $\mathbf{C} = \mathbf{I} - p^{-1}\mathbf{u}_p\mathbf{u}_p^t$ is a centering matrix. The cross-product matrices are vectorized $\tau_r = vec(\mathbf{T}_r), r = 1, 2, \cdots, m$ and organized in a $p^2 \times r$ matrix $\mathbf{Z} = [\tau_1, \tau_2, \cdots, \tau_m]$. Let $\mathbf{A} = \mathbf{N}^{-0.5}(\mathbf{Z}^t\mathbf{Z})\mathbf{N}^{-0.5}$ where $\mathbf{N}$ is a diagonal matrix whose elements are the diagonal elements of $\left(\mathbf{Z}^t\mathbf{Z}\right)$. The generic element $a_{r,v}$ of $\mathbf{A}$ is the vectorial correlation coefficient between $\mathbf{S}_r$ and $\mathbf{S}_v$. We have $a_{r,v} > 0$ because the partial distance matrices are Euclidean. The decomposition of $\mathbf{A}$ yields $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^t$ with $\mathbf{Q}^t\mathbf{Q} = \mathbf{I}$. Since $\mathbf{A}$ is non negative, then its largest eigenvalue is positive and there is a $m \times 1$ positive eigenvector $\mathbf{q}$ corresponding to it. Thus, the compromise cross-product matrix can be found using

$$\mathbf{T} = \sum_{r=1}^{m} \alpha_r \mathbf{T}_r \quad \text{where} \quad \alpha = \left(\mathbf{u}_m^t\mathbf{q}\right)^{-1}\mathbf{q}. \tag{5}$$

The related compromise distance matrix is $\mathbf{D} = \Psi\mathbf{u}_p^t + \mathbf{u}_p\Psi^t - 2\mathbf{T}$ with $\Psi = diag(\mathbf{T})$ This matrix is the basis of clustering and classification methods based on distance matrix such as hierarchical or partitioning around medoids methods.

4. Analysis of the regional disparities in Italy

The aim of this section is to study the disparities among the $p = 20$ regions Italian regions by using a system of $m = 159$ demographic, social, environmental and economic indicators. Our sources of data were Italy's 2000-2006 Community Support Framework ("Quadro Comunitario di Sostegno", or "QCS") `http://www.istat.it/ambiente/contesto/azioneB.html` and Regional Economic Accounts ("Conti Economici Regionali", or "CER"), `http://www.istat.it/dati/dataset/20051025_00/`.
We have applied the distance matrix $\mathbf{D}$ together with the PAM algorithm available in the R-package *cluster*. We choose this algorithm because does not impose a hierarchical solution, which is not necessarily present in regional data. In addition, PAM determines a leading region for each group (the cluster medoid) as the region whose average dissimilarity to all other regions in the same cluster is minimal. The resulting clusters are:

Cluster 1 {*Abruzzo*, *Basilicata*, **Calabria**, *Campania*, *Molise*, *Puglia*, *Sardegna*, *Sicilia*}
Cluster 2 {*Emilia Romagna*, *Friuli Venezia Giulia*, *Lazio*, *Liguria*, *Lombardia*, *Marche* *Piemonte*, **Toscana**, *Trentino Alto Adige*, *Umbria*, *Veneto*}
Cluster 3 {**Val d'Aosta**}

The number of clusters was selected based on the average silhouette width. The clustering reflects the simple approach based on GDP per capita PPS (purchasing power standards) data used by European Union for structural funds allocation. Cluster 1 groups together regions that are "Objective 1", that is, regions with GDP/PPS per capita under 75% of the EU average. Cluster 2 consists of the regions falling out of Objective 1. The outlier in Cluster 3 is the small northern region of Valle d'Aosta that is usually excluded from regional analysis or aggregated to Piemonte. The medoids, indicated in bold type, are a multidimensional version of the median. A possible interpretation is as follows: if the regions' position on the different issues were a monotonous function of their individual level on an overall transversal factor, then the multidimensional median position of the cluster would be held by the region with the median of that factor. If the above clustering structure is appropriate then regions considered as similar share similar properties; this allows us to analyze and model only a selected number of representative regions in each group. In particular, the region of Calabria and Toscana could be compared to reveal different patterns of development or convergences between the clusters they represent.

5. Conclusions and future research

One important feature of our algorithm is that the distance matrix at the basis of clustering is obtained as a reasonable synthesis of all the multiple time series describing the dynamic relationships between a set of variables. The method is appealing because of its easy of interpretation and good performance with short time series. The main drawback of the method is the course of dimensionality: when large data sets are averaged, the differences between units are expected to be significantly reduced so that only a global and macroscopic pattern can emerge from the mist of huge data repositories. In our experiment on Italian data, the fusion of a great number of indicators observed for different time intervals has showed the evolution of development disparities between two blocks of regions: objective 1 regions and non-objective 1 regions.
Future research will focus on system to compress the information contained in MTS (*e.g.* multiway principal component analsys) to prevent the decline in performance when large data sets are compacted. This decline is especially relevant in classification where average are largely used for computing distances at the univariate level.The agglomeration phase must be better studied to compare different weighting scheme. The spatial nature of the objects to be clustered (regions) should also to be taken into account.