COMSTA: 6622 Model 3G pp. 1–13 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Classification tree methods for panel data using wavelet-transformed time series

Xin Zhao [a], Stuart Barber [a,*], Charles C. Taylor [a], Zoka Milan [b]

[a] School of Mathematics, University of Leeds, UK
[b] King's College Hospital Trust, London, UK

ARTICLE INFO

Article history:
Received 5 July 2017
Received in revised form 21 May 2018
Accepted 25 May 2018
Available online xxxx

Keywords:
CART
MODWT
Panel data
Noise exclusion

ABSTRACT

Wavelet-transformed variables can have better classification performance for panel data than using variables on their original scale. Examples are provided showing the types of data where using a wavelet-based representation is likely to improve classification accuracy. Results show that in most cases wavelet-transformed data have better or similar classification accuracy to the original data, and only select genuinely useful explanatory variables. Use of wavelet-transformed data provides localized mean and difference variables which can be more effective than the original variables, provide a means of separating "signal" from "noise", and bring the opportunity for improved interpretation via the consideration of which resolution scales are the most informative. Panel data with multiple observations on each individual require some form of aggregation to classify at the individual level. Three different aggregation schemes are presented and compared using simulated data and real data gathered during liver transplantation. Methods based on aggregating individual level data before classification outperform methods which rely solely on the combining of time-point classifications.

1. Introduction

We often encounter data containing multiple time series variables for organizations or individuals that need classification, especially in areas such as economics, finance, marketing, medicine and biology. It can also be important to determine which of the time series are useful in performing the classification; interpreting this information can be highly useful in investigating the relationships between the variables and the class labels.

Our interest in this problem was motivated by data collected on patients undergoing liver transplant surgery. Each patient is classified into one of two groups, according to whether they did or did not use beta-blocker medication. During the operation, monitoring took place for several variables such as heart rate and systolic blood pressure, with data recorded once every heartbeat. However, equivalent problems arise in many different contexts.

Denote the data as $A_{n,k,t}$ for individual (or organization) $n = 1, 2, \ldots, N$, variable $k = 1, 2, \ldots, K$, and time $t = 1, 2, \ldots, T_n$, allowing the length of the time series to be different for each individual. Thus, for the nth individual, the data can be expressed as a $T_n \times K$ matrix

$$A_{n,\cdot,\cdot} = \begin{bmatrix} A_{n,1,1} & \cdots & A_{n,K,1} \\ A_{n,1,2} & \cdots & A_{n,K,2} \\ \vdots & \ddots & \vdots \\ A_{n,1,T_n} & \cdots & A_{n,K,T_n} \end{bmatrix}.$$

* Correspondence to: School of Mathematics, University of Leeds, Leeds LS2 9JT, UK.
E-mail address: S.Barber@leeds.ac.uk (S. Barber).

https://doi.org/10.1016/j.csda.2018.05.019
0167-9473/

Please cite this article in press as: Zhao X., et al., Classification tree methods for panel data using wavelet-transformed time series. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.05.019.

The full data can be written as a $\sum_{n=1}^{N} T_n \times K$ matrix A, where

$$A^T = \left[A_{1,\cdot,\cdot}^T \quad A_{2,\cdot,\cdot}^T \quad \cdots \quad A_{N,\cdot,\cdot}^T\right].$$

We refer to such explanatory data as panel (or longitudinal) data. The response variable Y is the group that each individual belongs to, which we write as a vector

$$y^T = \left[y_{1,\cdot}^T \quad y_{2,\cdot}^T \quad \cdots \quad y_{N,\cdot}^T\right],$$

where $y_{n,\cdot}$ has T_n identical values, defined by $(y_{n,1}, y_{n,2}, \ldots, y_{n,T_n})$.

Difficulties in analyzing such data sets include: (1) unequal values of T_n; (2) aggregating the panel data to provide classification for each individual; and (3) lack of independence between consecutive times.

Such data are generally subject to noise if they are collected or recorded by people or machines. Wavelet shrinkage (Donoho and Johnstone, 1994) is a popular denoising method, which is commonly used to smooth out random noise variation in signals, and we could use such methods in our application. However, even without a formal denoising step, wavelets are able to separate out "signal" from "noise", and we use this property to improve prediction performance. We shall also see that wavelets can pick out short term fluctuations in real data which can be exploited for classification when consecutive observations lack independence. Instead of using the standard decimated discrete wavelet transform (DWT), we use the maximal overlap discrete wavelet transform (MODWT; see, for example, Percival and Walden, 2000, ch. 5), as it is not constrained by time series length T_n and each time point is represented at all resolution levels of the MODWT. Equivalent translation-equivariant transforms are the non-decimated stationary wavelet transform (Nason and Silverman, 1995) and cycle-spinning (Coifman and Donoho, 1995).

For classification, we use the classification and regression tree (CART) method of Breiman et al. (1984). Using DWT (or MODWT) with CART (or other decision trees or random forests) in time series data has already been considered (Alickovic and Subasi, 2016; Gokgoz and Subasi, 2015; Upadhyaya and Mohanty, 2016), as have other classification methods (Maharaj and Alonso, 2007, 2014) but, to the best of our knowledge, until now the application to panel data is quite rare. Previous authors have directly converted the wavelet representation of panel data into cross-sectional data by using summaries such as energy, standard deviation, or entropy (Zhang et al., 2015; Upadhyaya and Mohanty, 2016).

However detecting when and how MODWT can help CART in classification accuracy and variable selection for panel data is important. Thus, in this paper, we use CART with original and wavelet-transformed variables to classify panel data. We introduce our methodology in Section 2, and apply it to simulated panel data experiments in Section 3 before analyzing our liver transplantation (LT) panel data in Section 4. Some concluding comments appear in Section 5.

2. Methodology

In this section, CART and the MODWT are introduced briefly. For more details, see Breiman et al. (1984) and Percival and Walden (2000), respectively. We then propose three methods to produce individual-level classifications from panel data, which can be applied to the original data, the wavelet-transformed data, or a combination of both.

2.1. Background

The goal of CART is to construct a model that predicts the value of a response variable by learning simple decision rules inferred from data features. In our case, we use classification trees since the response variable is categorical. The model built is structured as a tree with each node representing a split of the data in that node according to the value of a single variable. The aim is to use successive decision rules to split the data in a way which makes the subset of data at each terminal node (leaf node) as pure as possible, ideally with only one class. The tree represents, therefore, a classification rule using only those variables which are found to convey as much relevant information as possible. The tree construction process involves building and then pruning a tree. In the tree building process, we use optimization of the Gini index as our splitting criterion to choose the best explanatory variable to construct a decision at each node. Specifically, the decision rule using the selected variable at each node will determine a subset of the data which is as pure as possible, as measured by the Gini index. In the pruning process, we use ten-fold cross-validation to mitigate the tendency of CART to produce over-fitted models. Overall, then, CART is used to construct a classification rule which incorporates a variable selection stage as part of the construction process.

In the MODWT, we use the Haar scaling function ϕ and wavelet ψ, where

$$\phi(\tau) = \begin{cases} 1 & \tau \in [0, 1) \\ 0 & \text{else,} \end{cases} \qquad \text{and} \qquad \psi(\tau) = \begin{cases} 1 & \tau \in [0, 1/2) \\ -1 & \tau \in [1/2, 1) \\ 0 & \text{else.} \end{cases} \tag{1}$$

By using dilation and translation, we obtain the scaling function and wavelet at location l and resolution level j:

$$\phi_{j,l}(\tau) = 2^{j/2}\phi\left(2^j(\tau - l)\right) \qquad \text{and} \qquad \psi_{j,l}(\tau) = 2^{j/2}\psi\left(2^j(\tau - l)\right),$$

where $j = 0, 1, \ldots, J$, for $J = \lfloor \log_2 n \rfloor$, and $l = 0, 1, \ldots, n-1$. Note that $\phi_{j,l}$ and $\psi_{j,l}$ are compactly supported on $I_{j,l} = \left[2^{-j}l, 2^{-j}(l+1)\right)$. When $j = 0$, the scaling coefficients are actually the original time series values.