Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 35 (2008) 1301–1310

Expert Systems with Applications

www.elsevier.com/locate/eswa

Wavelet-based multiresolution analysis for data cleaning and its application to water quality management systems

Li He [a], Guo-He Huang [a,b,*], Guang-Ming Zeng [c], Hong-Wei Lu [a]

[a] *Environmental Systems Engineering Program, Faculty of Engineering, University of Regina, Regina, Sask, Canada S4S 0A2*
[b] *Chinese Research Academy of Environmental Science, North China Electric Power University, Beijing 100012-102206, China*
[c] *College of Environmental Engineering and Science, Hunan University, Changsha, Hunan 410082, China*

Abstract

Data cleaning techniques are useful for extracting desirable knowledge or interesting patterns from existing databases in engineering applications. The major problems of conventional techniques (e.g., Fourier Transformation Technique) are that they are (1) more appropriate in linear systems than nonlinear systems, and (2) stringently depend on state space functions. In this study a wavelet-based multiresolution analysis technique (WMAT) is proposed for reducing noises induced by complex uncertainty. The approach is applied to a river water quality simulation system for showing its practicability in data cleaning and parameter estimation. Clean data are prepared through running a Thomas' river water quality model and polluted data are synthesized by mixing clean data with white Gaussian noises. The results show that WMAT will not distort the clean data, and can effectively reduce the noise in the polluted data. The data denoised by WMAT are furthermore used for estimating the modeling parameters. It is also indicated that the parameters estimated with the denoised data through WMAT are much closer to real values than those (1) with polluted data through WMAT and (2) with data through Fourier analysis technique. It is thus recommended that the prepared data be used for estimating the modeling parameters until being cleaned with WMAT.

Keywords: Data cleaning; Wavelet; Multiresolution analysis technique; Parameter estimation; Water quality management

1. Introduction

With the increasing ease of generating, collecting and storing data, we are living in an expanding universe of too much data (Sőrensen & Janssens, 2003). Extracting useful information is a must from these abundant data. Data mining is a process of extracting desirable knowledge or interesting patterns from existing databases for special purposes. The process mainly covers six stages: data collection, data preprocessing, feature extraction, patterns recognition, data visualization, and results evaluation. Conventional data mining techniques involve decision trees, multicriteria analysis (Zeng & Trauth, 2005), artificial neural networks (Zeng et al., 2003), statistical analysis (Battista & Visini, 2006), Bayesian data analysis (He, Chan, Huang, & Zeng, 2006), etc. There have been a variety of fields such as marketing, management, health care and other areas of computer, astronomy, bioinformatics, high-energy physics, chemistry, and environmental management (Hong, Lin, & Wang, 2003; Babovic, Drécourt, Keijzer, & Friss Hansen, 2002; Liu & Shih, 2005; Kusiak, Dixon, & Shah, 2005; Ye, 2003; Huang, 2006; Yin, Estberg, Hallisey, & Cayan, 2007; Fisher, Wood, & Cheng, 2007).

The above study efforts were normally based on an assumption that the data to be mined should be reliable and accurate. However, the data arising from investigation, experiment, and simulation processes may be polluted by noise signals due to the subjective and/or objective errors (Li & Shue, 2004; Mu, 1996). For example, the experiment

* Corresponding author. Address: CRAES, North China Electric Power University, Beijing 100012-102206, China. Tel.: +1 391 146 8225; fax: +1 306 585 4855.
E-mail address: huang@iseis.org (G.-H. Huang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.08.009

errors may be resulted from measurement, reading, recording, and external conditions; the simulation errors might cover model uncertainty, parameter uncertainty, and computation errors. Since these noisy signals are probably to distort the results of the data mining, it is a must to remove them (that is, signal denoising) before using any original data.

Signals can be denoised through the application of a set of linear filters (Bell & Martin, 2004; Constable, 1978; Mu, 1996). However, one problem of these filters is that they are more appropriate in linear systems than nonlinear systems. Another problem is that they are dependent of state space functions. While in fact, most of signals are nonlinear and can hardly be represented by a special state space functions. In addition, Fourier analysis technique (FAT) is a classical tool for reducing noises, but it is only suitable for denoising data/signals containing steady noises. Due to the noises that are unsteady in real-world cases, its application is still limited. To overcome the problems of traditional denoising techniques, more sophisticated techniques such as wavelet-based multiresolution analysis technique (WMAT) has been proposed.

WMAT is useful for denoising multi-dimensional spatial/temporal signals containing steady/unsteady noises. It has been widely applied to engineering systems for patterns recognition and knowledge discovery (Avci, 2007; Ceylan & Özbay, 2007; Duport, Girel, & Chassery, 1996; Galal, 2002; Hobbs & Hepenstal, 1989; Hsieh & Kuo, 2008; Li & Shue, 2004; Lung, 2006, 2007; Mallat, 1989; Murtagh, Starck, & Bijaoui, 1995; Osowski & Nghia, 2002; Otazu & Pujol, 2006; Schutze, 2001; Sorzano, Ortiz, & López, 2006; Subasi, 2007; Starck & Murtagh, 1998; Tirtom, Engin, & Engin, 2008). Nevertheless, few of these studies were applied to water quality management systems, where the water quality monitoring data needs to be used for parameter estimation (Dohan & Whitfield, 1997a, 1997b).

Therefore, the objective of this study is to propose a wavelet-based multiresolution analysis technique (WMAT) for cleaning the polluted water quality monitoring data. The technique, together with the traditional Fourier analysis technique (FAT), will be applied to a numerical example to illustrate the performance of WMAT in data cleaning. In addition, the denoised and non-denoised data will be simultaneously applied to a water quality management system for dealing with parameter estimation issues.

2. Wavelet-based multiresolution analysis technique

2.1. One-dimensional continuous wavelet transform

Suppose $\Psi(t)$ belongs to a two-dimensional space $L^2(R)$, whose Fourier transform is $\hat{\Psi}(\bar{\omega})$. If $\hat{\Psi}(\omega)$ satisfies the permitted condition (also called completely reconstruction condition):

$$C_\Psi = \int_R \frac{|\hat{\Psi}(\omega)|^2}{|\omega|}\mathrm{d}\omega < \infty \tag{1}$$

we call $\Psi(t)$ a basic wavelet or mother wavelet. If the flex and translation transform is conducted for mother wavelet $\Psi(t)$, we can get

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right) \quad a, b \in R;\ a \neq 0 \tag{2}$$

and correspondingly, we call $\Psi_{a,b}(t)$ a wavelet series for all the a and b, where a is the flex factor (that is, the scaling in frequency range), b is the translation factor. For any function $f(t) \in L^2(R)$, its one-dimension continuous wavelet transform is

$$W_f(a,b) = \langle f, \Psi_{a,b}\rangle = |a|^2 \int_R f(t)\hat{\Psi}\left(\overline{\frac{t-b}{a}}\right)\mathrm{d}t \tag{3}$$

where, $W_f(a,b)$ is called the wavelet coefficient. The continuous wavelet coefficients can be used to reconstruct function $f(t)$ and its reconstruction equation is

$$f(t) = \frac{1}{C_\Psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{a^2}W_f(a,b)\Psi\left(\frac{t-b}{a}\right)\mathrm{d}a\,\mathrm{d}b \tag{4}$$

2.2. One-dimensional discrete wavelet transform

The discrete equations for flex parameter a and translation parameter b are in general restricted to only discrete values such that

$$a = a_0^j, \quad b = ka_0^jb_0 \tag{5}$$

where $j \in Z$, a_0 is a fixed extended step. So, the discretion wavelet function $\Psi_{j,k}(t)$ can be written by

$$\Psi_{j,k} = a_0^{-j/2}\Psi\left(\frac{t - ka_0^jb_0}{a_0^j}\right) = a_0^{-j/2}\Psi(a_0^{-j} - kb_0) \tag{6}$$

Correspondingly, the discrete wavelet transform coefficient can be expressed by

$$C_{j,k} = \int_{-\infty}^{\infty} f(t)\Psi_{j,k}(t)\,\mathrm{d}t = < f, \Psi_{j,k} > \tag{7}$$

Thus, the reconstruction equation of $f(t)$ is

$$f(t) = C\sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty} C_{j,k}\Psi_{j,k}(t) \tag{8}$$

where C is a constant independent of data.

2.3. Multiresolution analysis technique

Mallat (1989) proposed a concept of multiresolution analysis for constructing orthonormal wavelet basis and further illustrated the wavelet's multiresolution characteristics from the space aspect. In addition, a fast algorithm for orthonormal wavelet was put forward. The works demonstrated the functions of the wavelet theory in frequency analysis for various data signals, especially of mutational or unsteady ones. The concept of the multiresolution analysis consists of two stages: decomposition and reconstruction, which can be illustrated by a multiresolution