JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN ELECTRONICS AND COMMUNICATION ENGINEERING

OUTLIER REMOVAL USING MULTILEVEL WAVELET TRANSFORM TECHNIQUE

[1] MONA J. SAGAR, [2] S. N. SAMPAT, [3] S.B.PARMAR

[1]Department Of Electronics & Communication Engineering, Shantilal shah Engineering College, Bhavnagar, Gujarat, India
[2] Department of Electronics & Communication Engineering, G.P., Gandhinagar, Gujarat, India
[3] Department of Electronics & Communication Engineering, Shantilal shah Engineering college, Bhavnagar, Gujarat, India

sagar.mona@gamil.com,sanjay_sampat@yahoo.in,shailbapu@yahoo.com

ABSTRACT: *In statistics, an outlier is an observation that is numerically distant from the rest of the data. In many applications such as fraud Detection, Medical Treatments, Weather Prediction etc. detection and removal of outlier is very important. Wavelet transform is an advance mathematical tool which can localize different components of signal under consideration with variable resolution matched to its scale. Hence using multilevel and multi resolution features of suitable wavelet, we can remove outliers while preserving the rest of the signal.*

Keywords: wavelet, outliers, distribution, multilevel analysis, scale.

1. INTRODUCTION

Wavelet transform techniques provide multiscale analysis of the signal as a sum of orthogonal signals corresponding to different time scales. So, it is called time-scale analysis. It provides multilevel analysis to analyse the signal at different levels.

In present paper, at first, we present the basic idea of outliers in terms of it different definitions, applications, causes and different technique for removing them. In the next section, we take one case study of electrical load consumption, sampled minute by minute, over a 5-week period. We used sample of this signal for detecting and removing the outliers. Here we have used "Daubechies" wavelet up to level 5.

2. INTRODUCTION TO OUTLIERS

Grubbs defined an outlier as: An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. According to mathematics the outliers define as: A data point that is distinctly separate from the rest of the data [4].An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. [5].When we are sampling a data, some data points may be further away from the sample mean value. It can be identical due to systematic error or flaws in theory or it may be that some observations are far from the centre of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers are to be expected [3]. Outliers can occur by chance of any distribution, but they are often indicative either of measurement error or that the population has an improper distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations.

3. APPLICATION OF OUTLIERS

1. Fraud Detection (Credit card, telecommunications, criminal activity in e-Commerce)
2. Customized Marketing (high/low income buying habits)
3. Medical Treatments (unusual responses to various drugs)
4. Analysis of performance statistics (professional athletes)
5. Weather Prediction
6. Financial Applications (loan approval, stock tracking) [2]

4. CAUSES OF OUTLIERS

1. Poor data quality / contamination
2. Low quality measurements, malfunctioning equipment, manual error
3. Correct but exceptional data [2]

5. OUTLIER DETECTION APPROACHES

1. Statistical-Based Outlier Detection
2. Deviation-Based Outlier Detection
3. Distance-Based Outlier Detection [2]

6. WAVELET AS A TOOL FOR OUTLIER DETECTION AND REMOVAL

Wavelet is a waveform of limited duration that has an average value of zero. Wavelet analysis allows signal to analyse with different resolution match to its scale. It is better than Fourier transform and Short time Fourier transform. It is used to analyse aspects like trends, break points, discontinuity at higher derivatives and self –symmetry, compression or de noising of the signal without appreciable degradation.[1]

Continuous wavelet transform is defined as,

$$CWT_x^{\Psi}(\tau,s) = \Psi_x^{\Psi}(\tau,s) = \frac{1}{\sqrt{|s|}}\int x(t).\Psi^*\left(\frac{t-\tau}{s}\right)dt \quad \ldots\ldots\ldots\ldots\ldots\ldots(1.1)$$

Where,

$\tau = translation\ (location\ of\ window)$;

$s = scale$

$\Psi^*\left(\frac{t-\tau}{s}\right) = mother\ wavelet$[8]

CWT is operating at every scale to analyze the signal. During computation CWT is continuous in terms of shifting i.e. the analyzing wavelet is shifted smoothly over the entire domain of the analyzed function [1].

DWT is useful to reduce computational time of analysis and synthesis by using only subset of scales for the positions at which to make our calculations.

7. ADVANTAGES OF WAVELET

A. *Denoising of the Signal*

Approximate coefficients are containing low frequency components while detail coefficients contain the high frequency components. When, we discard all high frequency information, we lose many of the information of original signal's sharpest features. Denoising is done by using suitable approach called 'thresholding'. In this approach detail coefficients are discarded when it exceeds certain limit. Denoised signal is reconstructed by using both the coefficients.

B. *Detecting long term evaluation*

Wavelet analysis may be used to detect the overall tread of the signal. As the approximation level increases the trend becomes clearer. Trend represents the slowest part of the signal. In terms of wavelet analysis, as the scale increases, resolution decreases. So, it produces better estimate of unknown trend. In terms of frequency, successive approximations possess progressively less high frequency information. With the higher frequencies removed, what is left is the overall trend of the signal.

C. *Splitting signal components*

The wavelet transform is used to split the signal in terms of its detail and approximate coefficients. The approximate coefficients represent the outlines and the detail coefficients represent detailed information. The detail coefficients are used to notice high frequency components and approximate coefficients are used to notice low frequency components.

D. *Detecting discontinuities and breakdown points*

These analyses are used to know at what exact instance the signal change occurs i.e. site of change, type of change, amplitude of change and discontinuities. By using detail coefficients at different level we can identify that the measurement and state noise. So by using wavelet transform we are able to detect the break down points.

E. *Multiscale analysis*

Wavelet technique provides Multiscale analysis of the signal as a sum of orthogonal signals corresponding to different time scale. So it is time scale analysis.

F. *Compression*

The wavelet transform denoised the signal by applying appropriate thresholding rule. Thresholding means removing the coefficients which are responsible for noise. So because of reduction in coefficients the signal is compressed without any original signal degradation.

To avail the above advantages while applying wavelet transform on the signal under consideration, the chosen wavelet from the list of available wavelets or any user defined wavelet must satisfy certain properties as mentioned below.

8. PROPERTIES OF WAVELET

A. *Admissibility*

CWT of signal is expressed by equation 1.1. Now, to recover original signal inverse CWT can be exploited as,

$$x(t) = \int_0^{\infty}\int_{-\infty}^{\infty}\frac{1}{s^2}X_{\omega}(s,\tau)\frac{1}{\sqrt{|s|}}\tilde{\psi}\left(\frac{t-\tau}{s}\right)d\tau\,ds$$

$\tilde{\psi}(t)$ Is, dual function of $\psi(t)$, and dual function should satisfy,

$$\int_0^{\infty}\int_{-\infty}^{\infty}\frac{1}{|s^2|}\psi\left(\frac{t_1-\tau}{s}\right)\tilde{\psi}\left(\frac{t-\tau}{s}\right)d\tau\,ds = \delta(t_1-t).$$

Sometimes, $\tilde{\psi}(t) = C_{\psi}^{-1}\psi(t)$

Where, $C_{\psi} = \frac{1}{2}\int_{-\infty}^{\infty}\frac{|\hat{\psi}(\omega)|^2}{|\omega|}\,d\omega$

C_ψ *is coefficent*

is called the admissibility constant and $\ddot{\psi}$ is Fourier transform ofψ. For a successive inverse transform, the admissibility constant has to satisfy the admissibility condition.

$$0 < C_\psi < +\infty.$$

According to admissibility condition, the wavelet must oscillate such that it's mean value equal to zero [9].

B .Regularity

It is useful in order to obtain reconstructed smooth and regular signals or images.*[3].*The definition and concept of regularity is somewhat technical. The regularity s of the signal f is defined as, the signal is s-time continuously differentiable at x_0 and s is an integer ($\geq$ 0) then regularity is s. the greater s, the more regular the signal. The regularity of certain wavelets is known.

C. Orthonormality:

It belongs to the discrete wavelet transform. Orthonormality is required to preserve the energy during transformation.

D. Multiresolution wavelet analysis

Wavelet transform allows multiresolution analysis by varying scale factor. The main purpose of it is to analyze the signal in multiple frequency bands. So that, the signals in multiple frequency bands are analyze differently and independently. The wavelet is therefore localities the signal time and frequency domains. So, multiresolution allows the signal to divide into different sub bands of different frequency and each subband is analyzed with a resolution matched to the scales of the wavelets.

E. Symmetry:

It is useful in order to avoid dephasing. Daubechies wavelet is asymmetric. While Coiflets and Symlets are near symmetric. [10]. Dephsing is of concern while one is dealing with image analysis and synthesis.

9. CASE STUDY

For outlier detection we are taking the case study of "Electricity consumption of France": In this example we consider a minute per minute record of electrical consumption of France. This problem is presented in detail in [MIS 94].Following points are taken in to consideration:

1. The load curve is aggregation of hundreds of sensors measurements, thus generating measurement errors.
2. In this case study we consider that the 50% components of load curve are due to industry, which appears as a regular profile and exhibits low-frequency changes. The remaining components are because of consumption of individual consumers. That appears as highly irregular, leading to high frequency components.
3. There are more than 10 million individual consumers.
4. The fundamental periods are the weekly-daily cycles, linked to economic rhythms.
5. Daily consumption patterns also change according to rate changes at different times (e.g. relay-switched water heaters to benefit from special night rates).
6. Missing data have been replaced.
7. Outliers have not been corrected.
8. For the observations 2400 to 3400, the measurement errors are unusually high, due to sensors failures. [1]

Here we are using distance based outlier detection. Two methods may be used: automatic detection and manual detection. Here manual detection is used. The figure 1 below shows the original electric signal in the range from 1150 to 1240 and its forward detail coefficents up to level 5 in which outliers are detected. The detailed coefficients of the relevant levels are made zero in order to remove that outliers and then the reconstructed signal is shown in the figure 2. Here the advantage of multilevel wavelet analysis is that instead of making all detailed coefficients zero, one can select the relevant coefficients so that the desired signal value is not affected much.

In the figure the outliers are indicated at time t=1193 and t= 1215 which is shown as black circle. Effect of outliers is clearly notice in lower level of detail coefficents. As far as outliers in concern the detail coefficents at level 1 and 2 are synchronize with the signal and futher levels are delay with the signal as shown in figure 1.

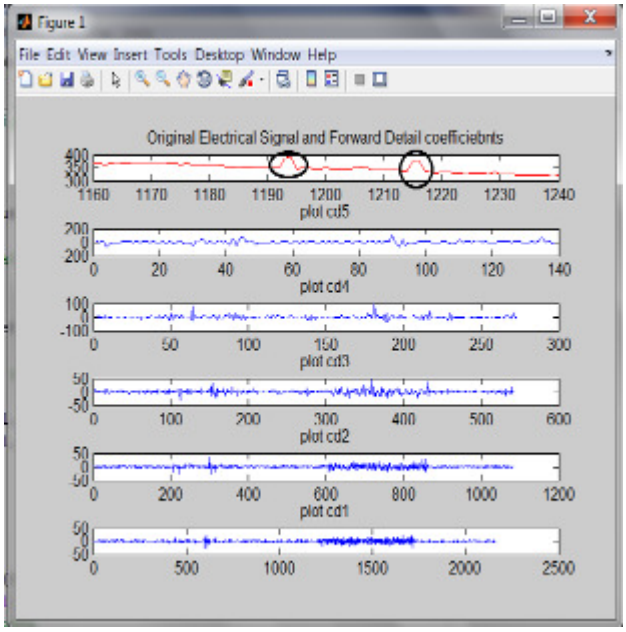

Fig 1. Original Electric Signal and Forward Detail Coefficients

Reconstruction means by use of approximate coefficients and modified detail coefficients we are able to reconstruct the signal. Here we reconstruct the

signal by suppressing the outliers by setting the corresponding values of the details to 0.

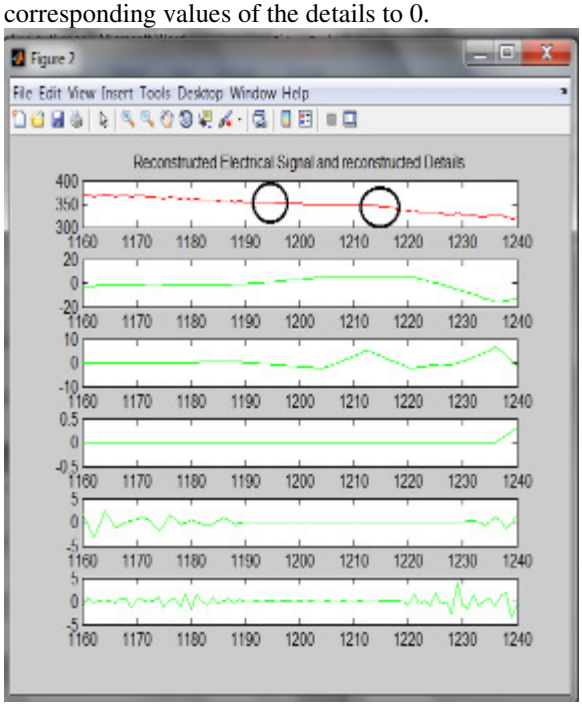

Fig 2. Reconstructed Signal and Reconstructed Detail Coefficients

When we are removing detail coefficients up to level 3. The same result is obtained without removing additional coefficients.

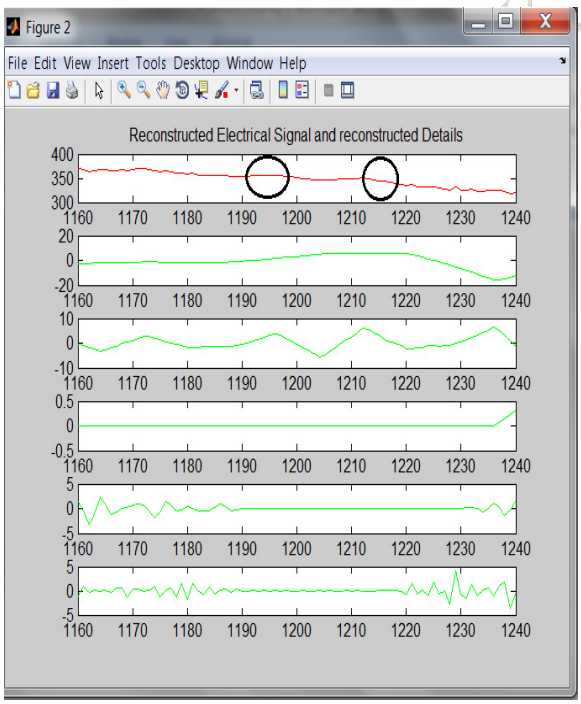

Fig 3. Reconstructed Signal and Reconstructed Detail Coefficients

In figure 3.after suppressing the detail coefficients up to level 3 relevant to outliers, we reconstruct the signal by using modified detail coefficients and approximate coefficients. Detail coefficients indicate high frequency information while approximate coefficients indicate low frequency information. Here we are not modified any of approximate coefficients. So as compare with figure 2 in figure 3 we reconstruct the signal with much loss of information. So we can get the same result by losing much less information.

10. TOOLS USED FOR CASE STUDY

Here, we have developed Matlab code to remove outliers.

11. CONCLUSION

If multilevel and multi resolution features of wavelet analysis at different scales are used appropriately, one can selectively remove unwanted outliers by proper selection of coefficients without losing desired part of the signal. So, multi-scale aspect is the most interesting and the most significant feature.

12.REFRENCES

[1] MATLAB 7.8.0. 347 (R2009a) Help Documentation
[2] Outlier Detection & Analysis By: Eric Poulin and Colin Yu
[3]http://en.wikipedia.org/w/index.php?title=Outlier&oldid=540653716
[4] Renze, John, "Outlier" from Math World
[5] Engineering Statistics Handbook
[6] *"Fundamental concept & an overview of the wavelet theory"* part 1 by ROBI POLIKAR.
[7] *"Fundamental concept & an overview of the wavelet theory"* part 2 by ROBI POLIKAR.
[8] *"Fundamental concept & an overview of the wavelet theory"* part 3 by ROBI POLIKAR
[9] Sheng, Y. "Wavelet Transform.", *The transforms and applications Handbook*: second edition, chapter 10 by Ed. Alexander D. Poularikas
[10] Wavelet and their application by Georges Oppenheim.