Procedia Computer Science

Volume 53, 2015, Pages 39–47

2015 INNS Conference on Big Data

Segmentation of geophysical data: a big data friendly approach

David R.B. Stockwell[1], Ligang Zhang[1], and and Brijesh Verma[1]

Centre for Intelligent and Networked Systems at Central Queensland University, Australia
(d.stockwell@cqu.edu.au, ligzhang@gmail.com, b.verma@cqu.edu.au)

Abstract
A new scalable segmentation algorithm is proposed in this paper for the forensic determination of level shifts in geophysical time series. While a number of segmentation algorithms exist, they are generally not 'big data friendly' due either to quadratic scaling of computation time in the length of the series N or subjective penalty parameters. The proposed algorithm is called SumSeg as it collects a table of potential break points via iterative ternary splits on the extreme values of the scaled partial sums of the data. It then filters the break points on their statistical significance and peak shape. Our algorithm is linear in N and logarithmic in the number of breaks B, while returning a flexible nested segmentation model that can be objectively evaluated using the area under the receiver operator curve (AUC). We demonstrate the comparative performance of SumSeg against three other algorithms. SumSeg is available as an R package from the development site at http://github.com/davids99us/anomaly.

Keywords: data segmentation, data size, geophysical, change points

1 Introduction

Has the level of a time series changed due to natural variation or an external influence? Abrupt changes in level can be due to instrument faults or reconfiguration and so are necessary for QA/QC on data from weather stations [1] and automatic tide or stream level gauges. The level changes in a segmentation model may also represent gene expression in micro-array comparative genomic hybridization data [2], regime shifts in climate data [3], breakouts in stock prices, twitter or web service logs, or features of interest in weak machine learning classifiers [4]. Finding an optimal multi-segmentation is challenging as the number of potential segments grows exponentially in N (as the number of potential ways to segment a sequence is equal to the number of subsets of N, or 2^N). Thus segmentation is an example of a big data problem where larger data sets call for new approaches [5]. The following are key performance criteria:

1. Linear or better order of increase in the computational cost of data length N and number of breaks B. Quadratic growth in computation cost does not scale.

2. Reliable application to noisy real world data with gaps, missing values and errors. Some algorithms fail on missing data, while others become unreliable or produce biased statistical measures.

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015

doi:10.1016/j.procs.2015.07.277

3. A well-known and transparent statistical framework. Type I/II errors (false positives and negatives) are unfortunately incorporated in only a few segmentation algorithms (CLAC and ACE) [2].

4. Objective model evaluation metrics like the Receiver Operating Characteristic (ROC) and optimization using the Area Under Curve (AUC).

5. Robust to non-normal data. Some geophysical data such as rainfall is highly non-normal, and most have strong periodic elements at a fine scale.

6. Flexible examination of models nested by confidence level. Output of a single model is not conducive to "drilling down" into sections with uncertainty.

This paper has three major contributions: 1) a novel ternary split segmentation algorithm available as a R package based on minimum and maximum extrema of the partial sums; 2) identification of linearity in length of data and number of breaks as crucial computational criteria for scaling segmentation; 3) use of the familiar learning statistical metric of the AUC as the criterion for breakpoints.

2 Related Work

Statistical methods for testing the homogeneity assumptions of linear models have a long history [6, 7]. An example of a simplistic algorithm would be to test each point for possible level changes using a goodness-if-fit test such as the standard normal homogeneity test (SNHT) [8]. As the goodness of fit of a segmented model increases without limit with additional breaks, additional penalties and measures such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) have been used to control over-fitting [9]. Combining the goodness of fit and penalties for over-fitting, the problem becomes a minimization of a global cost function:

$$\sum_{j=1}^{B} \mathcal{C}(y_{j:next(j)}) + \beta(B) \tag{1}$$

where $\mathcal{C}$ is a goodness of fit function for each segment and β is the penalty function usually over the number of breaks B. While the brute-force search for the optimum of equation 1 is $\mathcal{O}(2^N)$ dynamic programming can reduce the exact solution to $\mathcal{O}(BN^2)$ [10]. Polynomial growth in computation cost still imposes practical limits on N, and so is undesirable for big data analysis. Users face a confusing choice of the ideal fit and penalty functions so introducing potential operator bias (see [11]).

Segmentation algorithms are widely used in the detection and correction of meteorological data sets as station moves or reconfigurations often cause step changes in temperature. The International Surface Temperature Initiative (ISTI) is building global homogeneous temperature products [12] from a network of inhomogenous meteorological station data. Such projects need a very reliable segmentation method in the analysis chain. This is because segmentation is applied to a contrast with regional climatology, using either a weighted average of neighbors [13] or an exhaustive pairwise comparison [14]. While this reduces the noise from common climate variations, biases and errors may be introduced from the comparators [15]. Analysis of the original data collected at daily and shorter intervals is generally preferable as additional steps in the analysis chain such as monthly or annual aggregation can also introduce bias [16].

Here we compare the algorithm SumSeg with three representative approaches to multiple change-point detection algorithms supplied in the R package **changepoint** [11]. BinSeg, a binary segmentation algorithm, performs a recursive descent on binary splits blocking on segments