Segmenting Big Data Time Series Stream Data

Dima Alberg
SCE - Shamoon College of Engineering, Israel

Zohar Laslo
SCE - Shamoon College of Engineering, Israel

INTRODUCTION

Big data time series data streams are ubiquitous in finance, meteorology and engineering. It may be impossible to process an entire "big data" continuous data stream or to scan through it multiple times due to its tremendous volume. In Heraclitus's well-known saying, *"you never step in the same stream twice,"* and so it is with "big data" temporal data streams.

Unlike traditional data sets, big data continuous data streams flow into a computer system continuously, in a non-stationary way and with varying update rates. They are time-stamped, fast-changing, massive, and potentially infinite.

Under these circumstances, they represent an application area of growing importance in the data mining research. For example, sensors generate one million samples every minute (Hulten & Domingos, 2003) therefore the primary purpose of time series data stream segmentation is dimensionality reduction. This technique is used in many areas of data stream mining as: frequent patterns finding, structural changes and concept drifts detection (Ge & Smyth, 1999), time series classification and prediction (Hulten & Domingos, 2003), time series similarities searching (Keogh, Chakrabarti, Pazzani, & Mehrotra, 2000), (Park, Kim, & Chu, 2000), etc. The main principle of segmentation algorithms concludes in reducing the big data time series dimensionality by dividing the time axis into intervals behaving approximately according to a simple model. A good big data time series data stream segmentation algorithm must be OFASC (Online, Fast, Accurate, Simple and Comparable). For example the Sliding Window algorithm (Keogh, Chu, Hart, & Pazzani, 2004) on the one hand is online (O), very fast (F) and relatively simple (S) for using in online segmentation applications but on the other hand, it sometimes gives poor accuracy (A) and does not allow to perform online multivariate segmentation (C). Therefore, we will classify this algorithm to OFS segmentation algorithms domain.

The segmentation problem can be defined in following way: first, given a time series data stream to produce the best representation such that the maximum error for any segment does not exceed some user specified confidence level error threshold. It is important to add, that using a relative parameter such as confidence level will allow to evaluate an online multivariate segmentation and second, to construct a user friendly segmentation application which will evaluate and compare the proposed online segmentation algorithms in real time. As we shall see in later sections, the state-of-the-art segmentation algorithms do not meet all these requirements.

The rest of the paper is organized as follows. In Section 2, we provide a literature review of three state-of-the-art online piecewise linear segmentation algorithms. In Section 3, we provide a methodology for improving the existing state-of-the-art online segmentation algorithms. The proposed methodology based on Hoeffding bound error estimation, which uses a relative probability parameter instead of maximum error nominal parameter and meets the proposed OFASC re-

DOI: 10.4018/978-1-4666-5202-6.ch191

quirements. Section 4 briefly demonstrates a real-time segmentation application. Finally, in Section 5 and 6 we provide brief and meaningful empirical comparison of the proposed algorithms and suggest final conclusions.

BACKGROUND

Several high level representations of time series have been proposed in the research literature, including Fourier Transforms (Keogh et al., 2000), Wavelets (Chan & Fu, 1999), Symbolic Mappings (Das, Lin, Mannila, Renganathan, & Smyth, 1998; Perng et al., 2000) and Piecewise Linear Approximation or PLA: (Chan & Fu, 1999; Ge & Smyth, 1999; Hunter & McIntosh, 1998; Junker, Amft, Lukowicz, & Tröster, 2008; Keogh et al., 2004; Lavrenko, Schmill, Lawrie, Ogilvie, Jensen, & Allan, 2000; Li, Yu, & Castelli, 1998; Osaki, Shimada, & Uehara, 1999; Park, Lee, & Chu, 1999; Qu, Wang, & Wang, 1998; Shatkay & Zdonik, 1996; Vullings, Verhaegen, & Verbruggen, 1997; Wang & Wang, 2000).

In this work, our attention will confine to PLA, perhaps the most frequently used representation in continuous time series data streams. Obviously, all piecewise linear segmentation algorithms can also be classified as batch or online (Vullings et al., 1997). The problem discussed by (Keogh et al., 2004) is actually how to build online, fast and accurate algorithm for piecewise linear segmentation of time series data stream, because on the one hand, the main problem of online Sliding Window algorithm (Keogh et al., 2004) concerns in its poor accuracy (Qu et al., 1998; Wang & Wang, 2000) and its inability to look ahead. On the other hand the offline accurate Bottom Up (Keogh et al., 2004) algorithm is impractical or may even be unfeasible in a data mining context, where the data are in the order of terabytes or arrive in continuous streams. This problem is very important because for scalability purposes the proposed piecewise linear segmentation algorithm needs to capture the online nature of sliding windows and yet retain the superiority of Bottom Up.

In 2004 Keogh et al. (Keogh et al., 2004) introduced online Sliding Window Bottom Up (SWAB) algorithm which scales linearly with the size of the dataset, requires only constant space, produces high quality approximations of the initial time series data, and can be seen as operating on a continuum between the two extremes of Sliding Windows and Bottom-Up. The authors have shown that the most popular Sliding Window approach generally produces very poor results, and that while the second most popular approach, Top-Down, can produce reasonable results, it does not scale well with massive time series stream data.

MAIN FOCUS

As indicated in (Keogh et al., 2004), the main problem with the Sliding Windows algorithm is its inability to look ahead, lacking the global view of its offline (batch) counterparts. The Bottom-Up and the Top-Down (Junker et al., 2008; Keogh et al., 2004) approaches produce better results, but are offline and require the scanning of the entire data set. The SWAB algorithm has three nominal input parameters, which need to be defined carefully by the user in order to obtain an accurate segmentation model. For example, often the user obstructs to determine for the value of the maximal error threshold, because the data has very noisy non-stationary behavior. Therefore, in order to produce an accurate segmentation model, the user needs to perform the preprocessing of the obtained data or to perform a time consuming experiment design. Second, the inner loop of the SWAB algorithm simply invokes the Bottom-Up algorithm each time. This results in some computation redundancy and increases the computational complexity of algorithm. In this paper we introduce two new algorithms ISW (Improved Sliding Window) and ISWAB (Improved Sliding Window and Bottom Up) which decreases computational redundancy