Segmentation of Time Series Data

Parvathi Chundi
University of Nebraska at Omaha, USA

Daniel J. Rosenkrantz
University of Albany, SUNY, USA

INTRODUCTION

Time series data is usually generated by measuring and monitoring applications, and accounts for a large fraction of the data available for analysis purposes. A time series is typically a sequence of values that represent the state of a variable over time. Each value of the variable might be a simple value, or might have a composite structure, such as a vector of values. Time series data can be collected about natural phenomena, such as the amount of rainfall in a geographical region, or about a human activity, such as the number of shares of Google™ stock sold each day. Time series data is typically used for predicting future behavior from historical performance. However, a time series often needs further processing to discover the structure and properties of the recorded variable, thereby facilitating the understanding of past behavior and prediction of future behavior. Segmentation of a given time series is often used to compactly represent the time series (Gionis & Mannila, 2005), to reduce noise, and to serve as a high-level representation of the data (Das, Lin, Mannila, Renganathan & Smyth, 1998; Keogh & Kasetty, 2003). Data mining of a segmentation of a time series, rather than the original time series itself, has been used to facilitate discovering structure in the data, and finding various kinds of information, such as abrupt changes in the model underlying the time series (Duncan & Bryant, 1996; Keogh & Kasetty, 2003), event detection (Guralnik & Srivastava, 1999), etc.

The rest of this chapter is organized as follows. The section on **Background** gives an overview of the time series segmentation problem and solutions. This section is followed by a **Main Focus** section where details of the tasks involved in segmenting a given time series and a few sample applications are discussed. Then, the **Future Trends** section presents some of the current research trends in time series segmentation and the **Conclusion** section concludes the chapter. Several important terms and their definitions are also included at the end of the chapter.

BACKGROUND

A time series is simply a sequence of data points. In a segmentation of a given time series, one or more consecutive data points are combined into a single *segment* and represented by a single data point or a model for the data points in the segment. Given a time series *T* of *n* data points, segmentation of *T* results in a sequence of *m* segments, where each segment represents one or more consecutive data points in *T*. The number of segments *m* is typically much less than *n*. The input to a given instance of the segmentation problem is usually a time series and an upper bound on the number of segments.

Since the data points in a segment are represented by a single point or model, there is usually some error in the representation. Formally, the segmentation problem can be defined as follows. Given time series *T* and integer *m,* find a minimum error segmentation of *T* consisting of at most *m* segments. A specific version of this problem depends on the form of the data and how the segmentation error is defined. There are several approaches in the literature for addressing the segmentation problem. An *optimum* solution for the segmentation problem can be found by a dynamic programming based approach. (Duncan & Bryant, 1996; Gionis & Mannila, 2005; Himberg, Korpiaho, Mannila, Tikanmaki & Toivonen, 2001). A dynamic programming algorithm for solving this optimization problem runs in $O(n^2m)$ time, and so may not be practical for long time series with thousands of data points. There are more efficient heuristics that can be used to construct a segmentation of a given time series. However, these heuristics generally produce a suboptimal segmentation (Himberg, Korpiaho, Mannila, Tikanmaki & Toivonen, 2001; Keogh, Chu, Hart & Pazzani, 2004). There are also Bayesian based, fuzzy clustering, and genetic algorithm approaches to the segmentation problem (Oliver & Forbes, 1997; Abonyi, Feil, Nemeth, & Arva, 2005; Tseng, Chen, Chen & Hong, 2006). Methods have also been developed where a time series is segmented by converting it into a sequence of discrete symbols (Chung, Fu, Ng & Luk, 2004).

References (Chundi & Rosenkrantz, 2004a; Chundi & Rosenkrantz, 2004b) discuss segmentation algorithms for time series where each data point is a set of documents, each containing a set of keywords or key phrases. Reference (Siy, Chundi, Rosenkrantz & Subramaniam, 2007) gives an application of segmentation for time series where each data point is a set of items. References (Cohen, Heeringa & Adams, 2002; Gionis & Mannila, 2005) discuss segmentation algorithms for time series where each data point is a single symbol from an alphabet.

MAIN FOCUS

The main focus of a segmentation algorithm is to find the best segmentation to represent the given time series. There are *two primary* tasks in the segmentation process: constructing the representation of a given segment, and constructing the overall segmentation.

Representation of a Given Segment

The data representation of a given segment is computed from the time series data points in the segment, and can be viewed as representing a model of a process that may have generated these data points. The representation of a segment depends on the type of the data points in the time series, and the kind of model to be used. The usual goal in constructing the representation of a given segment is to minimize the error between the representation and the data points in the segment. Each data point in a time series can be a single numeric value (Keogh & Kasetty, 2003; Keogh, Chu, Hart & Pazzani, 2004), a vector of numeric values (McCue & Hunter, 2004), a set of items (Siy, Chundi, Rosenkrantz & Subramaniam, 2007) , a set of documents (Chundi & Rosenkrantz, 2004a), or some other type of value.

When each time series data point is a single numeric value, the model underlying a segment is typically a curve fitted to the points in the segment. This curve is constructed from the data points in the segment, usually with the goal of minimizing the error between the sequence of values of the data points in the segment, and the sequence of values implied by the segment representation. This segment error is usually computed by combining the difference between the value of each data point and the value assigned to that time point by the segment representation. These differences are often combined using an L_p metrics. L_1 is the Manhattan distance, i.e., the sum of the magnitude of the differences. L_2 is the Euclidean distance, based on the sum of the squares of the differences. In addition to the L_p metrics, local PCA methods have been used to measure error (Abonyi, Feil, Nemeth & Arva, 2005).

The curve for a given segment may be a single number (piecewise constant representation), a linear function (piecewise linear representation), a quadratic function, or a polynomial of even higher degree. A piecewise constant representation is a single numeric value that minimizes the segment error. For the L_1 error metric, this value is the median of the data points in the segment, and for the L_2 metric, it is the average of the data points (Gionis & Mannila, 2005). A piecewise linear representation fits a straight line to the data points (Keogh & Kasetty, 2003), usually with the goal of minimizing the L_2 error metric (Edwards, 1976). A piecewise constant or piecewise linear representation can be computed more efficiently than a quadratic or higher degree representation (Lamire, 2007). Sometimes the segments in a segmentation have different types of representation; e.g., some segments may have a linear representation, and some may have a quadratic representation, etc. (Lamire, 2007; Pednault, 1991).

When each data point is a vector of numeric values, a segment may be represented as a vector of constant values (McCue & Hunter, 2004), a vector of line segments (Abonyi, Feil, Nemeth & Arva, 2005), a vector of quadratics, etc.

When each data point is a set of documents, each of which contains a set of keywords or key phrases, the segment error is a measure of how closely the keywords (or key phrases) for the segment correspond to those for the documents in the time points of the segment (Chundi & Rosenkrantz, 2004a; Chundi & Rosenkrantz, 2004b).