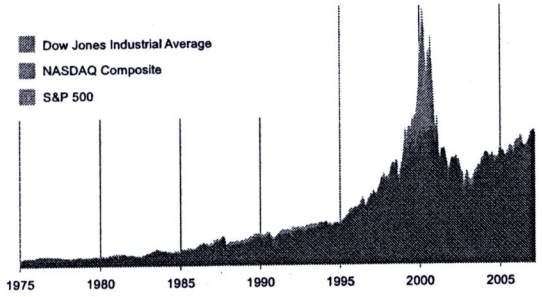
CHAPTER I

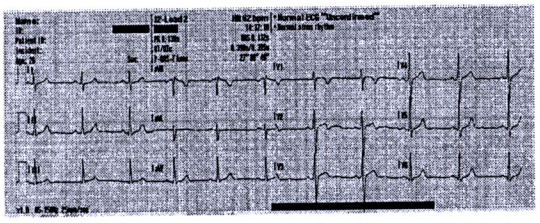
INTRODUCTION

Time series data mining is an active research area which involves tasks including classification (Ueno et al., 2006; Kasetty et al., 2008; Ratanamahatana and Keogh, 2004; Niennattrakul and Ratanamahatana, 2007c), clustering (Lin et al., 2004b; Yankov and Keogh, 2006; Niennattrakul and Ratanamahatana, 2007b, 2006), anomaly detection (Keogh et al., 2005, 2002; Yankov et al., 2008a; Niennattrakul et al., 2010a), pattern discovery (Chiu et al., 2003; Yankov et al., 2007; Mueen et al., 2009), visualization (Lin et al., 2004a; Kumar et al., 2005), association rules (Sacchi et al., 2007; Wan et al., 2007), and indexing (Keogh et al., 2004; Keogh and Ratanamahatana, 2005; Shieh and Keogh, 2009; Niennattrakul et al., 2010b). Time series is a sequence of real/integer/symbolic values which are sequentially observed, where in some applications, a time series sequence is also considered to be a very high dimensional data object, where the number of dimensions is equal to the length of time series. A characteristic that makes time series differ from other data types is that adjacent dimensions are extremely related; the order of each dimension cannot be swapped. Time series is ubiquitous, where it is easily found in daily life such as stock market, electrocardiogram, and a temperature record, as shown in Figure 1.1. Normally, time series can be collected from scientific measurements such as a star light curve (Protopapas et al., 2005), respiration (Keogh et al., 2005), and winding (B.L.R., 2010). In addition, a 2-D image can be transformed to be time series by sequentially measuring distances from the centroid of an image to the edge (Ye and Keogh, 2011; Yankov et al., 2008b). Therefore, instead of image recognition in 2-D images, time series mining will require much less complexity. A video can also be transformed to a time series sequence by tracking a coordinate of a point of interest. Time series can be multivariate, which at the specific time, many channels from different sources are observed. For example, SmartCane (Wu et al., 2008), a device attached with many types of sensors to help doctors monitor the walk of elderly people (see Figure 1.2), has eight channels of data from two pressure sensors, a three-axis accelerometer, and three single-axis gyros. Data from motion capture (Cai and Ng, 2004) are also considered that each dimension is collected from movement of each sensor.

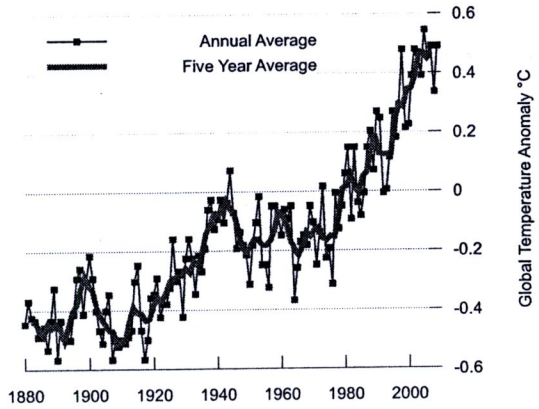
Subsequence clustering for time series data streams is an important data mining task which can return time series patterns in real time. Currently, no streaming subsequence clustering has yet been proposed. As a subsequence clustering result, cluster representatives can then be used in rule discovery (Das et al., 1998; Fu et al., 2001; Harms et al., 2002b,a; Hetland, 2002; Jin et al.,



a) Stock market (Wikipedia.org, 2011b)



b) Electrocardiogram (Wikipedia.org, 2011a)



c) Temperature record (Wikipedia.org, 2011c)

Figure 1.1: Examples of time series data in real world.

2002b,a; Mori and Kuni, 2001; Osaki et al., 2000; Sarker et al., 2003; Uehara and Shimada, 2002; Yairi et al., 2001), indexing (Li et al., 1998; Radhakrishnan et al., 2000), classification (Cotofrei, 2002; Cotofrei and Stoffel, 2002), prediction (Schittenkopf et al., 2000), and anomaly detection (Yairi et al., 2001). However, the current subsequence clustering, Subsequence Time Series Clustering (STSC), has been proved both theoretically and empirically to produces meaningless results, i.e., sine waves regardless of input sequences. Figure 1.3 illustrates cluster representatives from STSC. Therefore, hundreds of works that use STSC as a preprocessing step and a subroutine also produces meaningless results. The causes of meaninglessness are twofold: inappropriate uses of Euclidean distance measure and Amplitude Averaging function. In other words, Euclidean distance and Amplitude Averaging cannot handle trivial-matched subsequences which are a set of contiguous subsequences that are very similar but have shifts in time domain since Euclidean distance and Amplitude Averaging compute dissimilarity and an averaged result in one-to-one manner. Figure 1.4 provides some examples of trivial-matched subsequences.

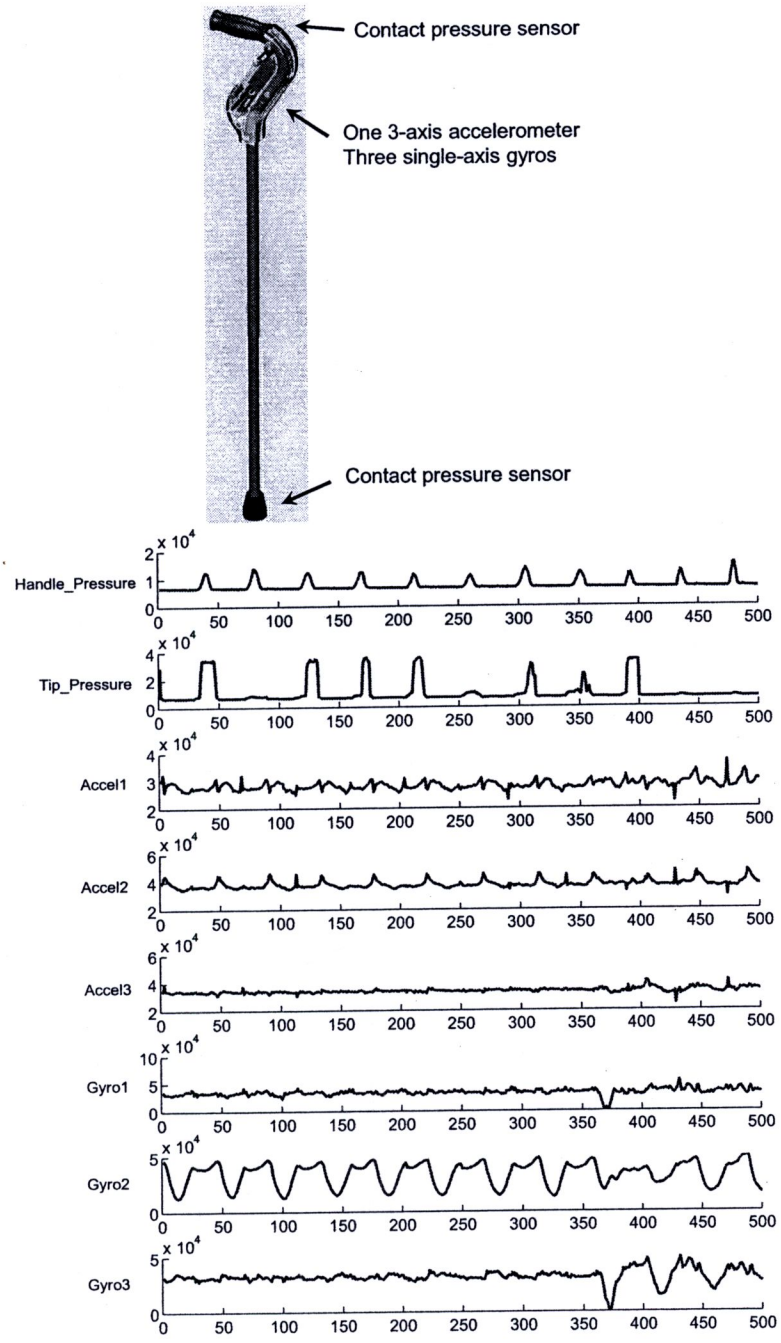


Figure 1.2: Multivariate time seires collected from SmartCane system. (Wu et al., 2008)

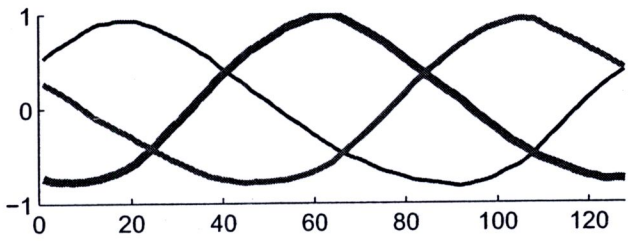
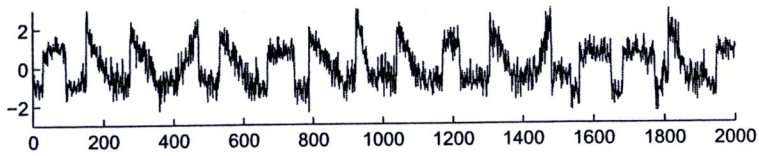
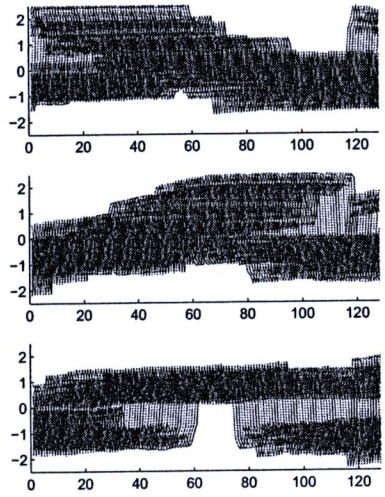


Figure 1.3: Cluster representatives generated from STSC



a) Some part of CBF sequence



b) Three sets of trivial-matched subsequences

Figure 1.4: Trivial-matched subsequences of CBF sequence

Recently, many researchers (Keogh and Lin, 2005; Denton, 2005; Chen, 2007a; Goldin et al., 2006; Fu et al., 2005; Struzik, 2003; Simon et al., 2006; Kumar et al., 2006; Fujimaki et al., 2008) attempt to overcome this problem by proposing many solutions. However, none of them propose the right solutions to deal with trivial-matched subsequences, i.e., new distance measures requires additional parameters and Amplitude Averaging is still used to create a cluster representative. The distance threshold in Density-based Subsequence Time Series Clustering (DSTSC) (Denton, 2005), the lag value in Lag-based Subsequence Time Series Clustering (LSTSC) (Simon et al., 2006; Chen, 2007b,a), and the slide length in Phrase-Analysis Subsequence Time Series Clustering (PASTSC) (Fujimaki et al., 2008), are additional parameters that users must be specified a priori, depending on characteristics of each dataset, whose values are very sensitive to clustering results. With incorrect values, outputs of clustering results may be meaningless. In addition, these values are used to discard trivial-matched subsequences; therefore, some important trivial-matched subsequences are unexpectedly filtered out. For the meaningfulness measurement, all previous works used Keogh-Lin Meaningfulness Measurement (KLMM) (Keogh et al., 2003) to measure clustering output. However, it will be demonstrated in this work that KLMM is an invalid measurement since it cannot capture similarity of sine waves with different phases and frequencies.

In this work, a novel subsequence clustering for data streams, Shape-based Streaming Sub-

sequence Time Series Clustering (3STSC), is proposed to return a meaningful clustering result in real time. Since all existing subsequence clustering algorithms produce meaningless result, to make subsequence clustering for data streams meaningful, subsequence clustering that produces meaningful results must first be introduced. In this work, a novel subsequence clustering for Shape-based Subsequence Time Series Clustering (2STSC), is firstly proposed. To produce meaningful clustering results, 2STSC utilizes Dynamic Time Warping (DTW) distance and Shape-based Averaging as a distance measure and an averaging function to replace Euclidean distance and Amplitude Averaging, respectively. DTW distance aligns subsequences before distance calculation; therefore, two trivial-matched subsequences are recognized as similar, and Shape-based Averaging aligns subsequences before averaging; therefore, a characteristic-preserved averaging result are returned from two trivial-matched subsequences. 2STSC is evaluated in terms of meaningfulness and this 2STSC is then extended to handle streaming cases in 3STSC.

The remaining of this dissertation is organized as follows. The meaninglessness of Subsequence Time Series Clustering (STSC) with the causes are analyzed and identified in Chapter II. Shape-based Averaging is first introduced in Chapter III. The solution to make a clustering result meaningful by Shape-based Subsequence Time Series Clustering (2STSC) is described and evaluated in Chapter IV. Incremental Shape-based Averaging is then proposed to extend Shaped-based Averaging to support streaming applications in Chapter V. Chapter VI provides a streaming subsequence clustering algorithm, Shape-based Streaming Subsequence Time Series Clustering (3STSC), which is extended from 2STSC to support streaming applications. And finally, this dissertation is concluded in Chapter VII.

1.1 Objective of the Thesis

The objective of this thesis is to design a novel subsequence clustering algorithm which produces meaningful clustering results for time series data streams.

1.2 Scopes of the Thesis

The scopes of this thesis are as follows:

- This thesis focuses on subsequence clustering for time series data streams, where the stream is univariate and a new data point arrives at a constant rate.
- The datasets from the Time Series Data Mining Archive (TSDMA) are used as benchmarks to evaluate subsequence clustering and streaming clustering, and the datasets from Time Series Clustering/Classification Page are used as benchmarks to evaluate shape-based aver-

aging and incremental shape-based averaging.

- Performance measurements used to evaluate the meaningfulness of the subsequence clustering algorithm is the Shape-based Meaningfulness Measurement (SMM), and the streaming subsequence clustering is evaluated by an actual time improved from the subsequence clustering.

1.3 Contributions of the Thesis

The contributions of this thesis are as follows:

- A new meaningfulness measurement is introduced.
- A novel subsequence clustering and a novel streaming subsequence clustering are proposed.
- A novel shape-based averaging and a novel incremental shape-based averaging are introduced.

1.4 Research Methodology

- Study background knowledge about time series data mining.
- Survey on potential and related topics including clustering, classification, anomaly detection, indexing, motif discovery, and subsequence matching.
- Review literatures on subsequence clustering algorithm.
- Identify causes of meaninglessness of the current subsequence clustering algorithm.
- Design the shape-based averaging algorithm as a major subroutine of subsequence clustering algorithm to solve the meaninglessness, and evaluate the algorithms with the benchmark datasets.
- Design the shape-based subsequence clustering algorithm that utilizes shape-based averaging algorithm to return a meaningful clustering result, and evaluate the algorithms with the benchmark datasets.
- Design the incremental shape-based averaging algorithm extended from shape-based averaging algorithm to support a streaming application, and evaluate the algorithms with the benchmark datasets.
- Design the shape-based streaming subsequence clustering algorithm extended from shape-based subsequence clustering algorithm to support a streaming application, and evaluate the algorithms with the benchmark datasets.

- Compose the thesis.