Clustering of Time Series Data

Anne Denton
North Dakota State University, USA

INTRODUCTION

Time series data is of interest to most science and engineering disciplines and analysis techniques have been developed for hundreds of years. There have, however, in recent years been new developments in data mining techniques, such as frequent pattern mining, that take a different perspective of data. Traditional techniques were not meant for such pattern-oriented approaches. There is, as a result, a significant need for research that extends traditional time-series analysis, in particular clustering, to the requirements of the new data mining algorithms.

BACKGROUND

Time series clustering is an important component in the application of data mining techniques to time series data (Roddick, Spiliopoulou, 2002) and is founded on the following research areas:

- **Data Mining:** Besides the traditional topics of classification and clustering, data mining addresses new goals, such as frequent pattern mining, association rule mining, outlier analysis, and data exploration (Tan, Steinbach, and Kumar 2006).
- **Time Series Data:** Traditional goals include forecasting, trend analysis, pattern recognition, filter design, compression, Fourier analysis, and chaotic time series analysis. More recently frequent pattern techniques, indexing, clustering, classification, and outlier analysis have gained in importance.
- **Clustering:** Data partitioning techniques such as k-means have the goal of identifying objects that are representative of the entire data set. Density-based clustering techniques rather focus on a description of clusters, and some algorithms identify the most common object. Hierarchical techniques define clusters at multiple levels of granularity. A survey of clustering that includes its application to time series data is provided in (Gan, Ma, and Wu, 2007).
- **Data Streams:** Many applications, such as communication networks, produce a stream of data (Muthukrishnan, 2003). For real-valued attributes such a stream is amenable to time series data mining techniques.

Time series clustering draws from all of these areas. It builds on a wide range of clustering techniques that have been developed for other data, and adapts them while critically assessing their limitations in the time series setting.

MAIN THRUST OF THE CHAPTER

Many specialized tasks have been defined on time series data. This chapter addresses one of the most universal data mining tasks, clustering, and highlights the special aspects of applying clustering to time series data. Clustering techniques overlap with frequent pattern mining techniques, since both try to identify typical representatives.

Clustering Time Series

Clustering of any kind of data requires the definition of a similarity or distance measure. A time series of length n can be viewed as a vector in an n-dimensional vector space. One of the best-known distance measures, Euclidean distance, is frequently used in time series clustering. The Euclidean distance measure is a special case of an L_p norm. L_p norms may fail to capture similarity well when being applied to raw time series data because differences in the average value and average derivative affect the total distance. The problem is typically addressed by subtracting the mean and dividing the resulting vector by its L_2 norm, or by working with normalized derivatives of the data

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

(Gavrilov et al., 2000). Several specialized distance measures have been used for time series clustering, such as dynamic time warping, DTW (Berndt and Clifford 1996), longest common subsequence similarity, LCSS (Vlachos, Gunopulos, and Kollios, 2002), and a distance measure based on well-separated geometric sets (Bollabas, Das, Gunopulos, and Mannila 1997).

Some special time series are of interest. A strict white noise time series is a real-valued sequence with values $X_t = e_t$ where e_t is Gaussian distributed random variable. A random walk time series satisfies $X_t - X_{t-1} = e_t$ where e_t is defined as before.

Time series clustering can be performed on whole sequences or on subsequences. For clustering of whole sequences, high dimensionality is often a problem. Dimensionality reduction may be achieved through Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), and Principal Component Analysis (PCA), as some of the most commonly used techniques. DFT (Agrawal, Falutsos, and Swami, 1993) and DWT have the goal of eliminating high-frequency components that are typically due to noise. Specialized models have been introduced that ignore some information in a targeted way (Jin, Lu, and Shi 2002). Others are based on models for specific data such as socioeconomic data (Kalpakis, Gada, and Puttagunta, 2001).

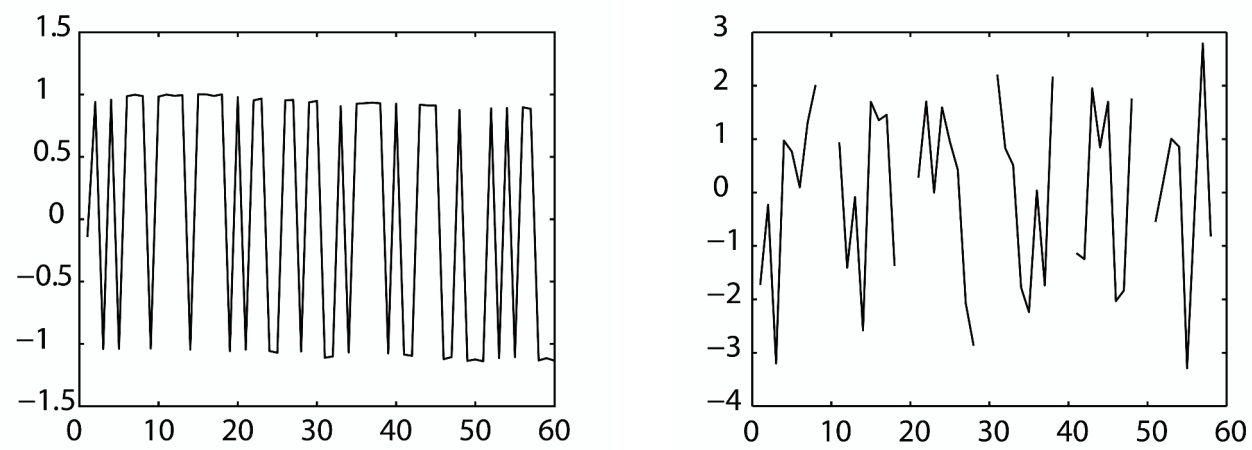
A large number of clustering techniques have been developed, and for a variety of purposes (Halkidi, Batistakis, and Vazirgiannis, 2001). Partition-based techniques are among the most commonly used ones for time series data. The k-means algorithm, which is based on a greedy search, has recently been generalized to a wide range of distance measures (Banerjee et al., 2004).

Clustering Subsequences of a Time Series

A variety of data mining tasks require clustering of subsequences of one or more time series as preprocessing step, such as Association Rule Mining (Das et al., 1998), outlier analysis and classification. Partition-based clustering techniques have been used for this purpose in analogy to vector quantization (Gersho and Gray, 1992) that has been developed for signal compression. It has, however, been shown that when a large number of subsequences are clustered, the resulting cluster centers are very similar for different time series (Keogh, Lin, and Truppel, 2003). Figure 1 illustrates how k-means clustering may find cluster centers that do not represent any part of the actual time series. Note how the short sequences in the right panel (cluster centers) show patterns that do not occur in the time series (left panel) which only has two possible values. A mathematical derivation why cluster centers in k-means are expected to be sinusoidal in certain limits has been provided for k-means clustering (Ide, 2006).

Several solutions have been proposed, which address different properties of time series subsequence data, including trivial matches that were observed in

Figure 1. Time series glassfurnace (left) and six cluster centers that result from k-means clustering with k=6 and window size w=8.



4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/clustering-time-series-data/10830

Related Content

Data Driven vs. Metric Driven Data Warehouse Design
John M. Artz (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 382-387).*
www.irma-international.org/chapter/data-driven-metric-driven-data/10848

Mining Repetitive Patterns in Multimedia Data
Junsong Yuan (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1287-1291).*
www.irma-international.org/chapter/mining-repetitive-patterns-multimedia-data/10988

Mining the Internet for Concepts
Ramon F. Brenaand Ana Maguitman (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1310-1315).*
www.irma-international.org/chapter/mining-internet-concepts/10991

Mining Data with Group Theoretical Means
Gabriele Kern-Isberner (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 1257-1261).*
www.irma-international.org/chapter/mining-data-group-theoretical-means/10983

Web Design Based on User Browsing Patterns
Yinghui Yang (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 2074-2079).*
www.irma-international.org/chapter/web-design-based-user-browsing/11105