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, which 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.
- **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 arbitrary levels of granularity.
- **Data Streams:** Many applications, such as communication networks, produce a stream of data. 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

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 (Gavrilov et al., 2000). Several specialized distance measures have been used for time series clustering, such as dynamic time warping, DTW (Berndt & Clifford 1996), and longest common subsequence similarity, LCSS (Vlachos, Gunopulos, & Kollios, 2002).

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, Faloutsos, & Swami, 1993) and DWT have the goal of eliminating high-frequency components that are typically due to noise. Specialized models have been
Related Content

Strategic Utilization of Data Mining
www.igi-global.com/chapter/strategic-utilization-data-mining/7724?camid=4v1a

A Hybrid Approach for Data Warehouse View Selection
www.igi-global.com/chapter/hybrid-approach-data-warehouse-view/7823?camid=4v1a

Online Signature Recognition
www.igi-global.com/chapter/online-signature-recognition/11012?camid=4v1a

Biomedical Data Mining Using RBF Neural Networks
www.igi-global.com/chapter/biomedical-data-mining-using-rbf/10575?camid=4v1a