Chapter II

Time Series Mining: Background and Related Work

co-authored with

Minghua Zhang, National University of Singapore, Singapore

A time series is a sequence of real numbers over time. Time series data occurs nearly everywhere. For example, the closing price of stock A over a week could be represented by a time series of \(\{23.80, 22.95, 21.87, 22.45, 23.32\}\). A time series database contains a set of time series. Current work on time series databases mainly focuses on indexing and querying, clustering and classification.

In this chapter, we will first give the background and review existing works in time series mining. The background material will include commonly used similarity measures and techniques for dimension reduction and data discretization. Then we will examine techniques to discover periodic and sequential patterns. This will lay the groundwork for the subsequent three chapters on mining dense periodic patterns, incremental sequence mining, and mining progressive patterns.
Issues in Time Series Mining

Similarity Measures

Time series mining problems are based on similarity computation, that is, to determine some numerical number that represents how similar two time series are. Therefore, it is important to choose a suitable similarity measure. Two similarity measures are commonly used: one is Euclidean distance, and the other is dynamic time warping (DTW).

Euclidean Distance

Euclidean distance is a simple yet widely used similarity measure. Given two time series $ts_1$ and $ts_2$ of length $L$, where

$$ts_1 = \{d_{11}, d_{12}, \ldots, d_{1L}\} \text{ and } ts_2 = \{d_{21}, d_{22}, \ldots, d_{2L}\},$$

the Euclidean distance between $ts_1$ and $ts_2$ is calculated as

$$d = \left[ (d_{11}-d_{21})^2 + (d_{12}-d_{22})^2 + \ldots + (d_{1L}-d_{2L})^2 \right]^{1/2}$$

The smaller the value of $d$, the closer is $ts_1$ and $ts_2$.

The Euclidean distance measure is easy to understand and can be efficiently computed. It has a time complexity of $O(n)$. However, in cases where phase shift in time axis exists, the Euclidean distance is not a suitable measure. For example, consider the following two time series

$$ts_1 = \sin(t)$$
$$ts_2 = \sin(t+10)$$

These two time series have similar shapes and differ only in the phase. Unfortunately, if we were to utilize the Euclidean distance as the similarity measure, we would conclude that $ts_1$ and $ts_2$ are quite different. This leads to the design of another similarity measure dynamic time warping (DTW).