Time Series Data Analysis

1. Introduction

A time series represents a collection of values obtained from sequential measurements over time. Time series data mining stems from the desire to rectify our natural ability to visualize the shape of data. Humans rely on complex schemes in order to perform such tasks. Major time series related tasks include query by content, anomaly detection, motif discovery, prediction, clustering , classification and segmentation.

Time series data mining unveils numerous facets of complexity. The most prominent problems arise from the high dimensionality of time series data and the difficulty of defining a form of similarity measure based on human perception. With the rapid growth of digital sources of information,
time series mining algorithms will have to match increasingly massive data-sets. The three major issues are involved:
— **Data representation**: How can the fundamental shape characteristics of a time series be represented?
— **Similarity measurement**: How can any pair of time series be distinguished or matched? This measure should establish a notion of similarity based on perceptual criteria, thus allowing the recognition of perceptually similar objects even though they are not mathematically identical.
— **Indexing method**: What indexing mechanism should be applied? The indexing technique should provide minimal space consumption and computational complexity.

2.Tasks in Time series data mining

2.1 Query by Content

Query by content is the most active area of research in time-series analysis. It is based on retrieving a set of solutions that are most similar to a query provided by the user.

Definition 2.1: Given a query time series Q = (q1, . . . , qn) and a similarity measure D(Q, T), find the ordered list L = {T1, . . . , Tn} of time series in the database DB, such that $\forall$ Tk, T j $\in$ L, k > j $\Leftrightarrow$D(Q, Tk) > D(Q, T j). The content of the result set depends on the type of query performed over the database.

Definition 2.2: (-Range Query). Given a query time-series Q = (q1, . . . , qn), a time-series database DB, a similarity measure D(Q, T) and a threshold , find the set of series S = {Ti | Ti $\in$ DB} that are within distance from Q. More precisely, find S = {Ti $\in$ DB | D(Q, Ti) $\leq$ }.

Definition 2.3 (K-Nearest Neighbors). Given a query time series Q = (q1, . . . , qn), a time-series database DB, a similarity measure D(Q, T), and an integer K, find the set of K series that are the most similar to Q. More precisely, find S = {Ti | Ti $\in$ DB} such that |S| = K and $\forall$ T j $\in$/ S, D(Q, Ti) $\leq$ D(Q, T j).
Such queries can be called on complete time series; however, the user may also be interested in finding every sub-sequence of the series matching the query, thus making a distinction between whole series matching and sub sequence matching. This distinction between these types of queries is thus expressed in terms of -range query.

2.2 Clustering

Clustering is the process of finding natural groups, called clusters, in a data-set. The objective is to find the most homogeneous clusters that are as distinct as possible from other clusters. More formally, the grouping should maximize intercluster variance while minimizing intracluster variance. The algorithm should thus automatically locate which groups are intrinsically present in the data.

The time-series clustering task can be divided into two subtasks.

2.2.1. Whole Series Clustering. Clustering can be applied to each complete time series in a set. The goal is thus to regroup entire time series into clusters so that the time series are as similar to each other as possible within each cluster.

Definition 2.2.1: Given a time-series database DB and a similarity measure D(Q, T), find the set of clusters C = {ci} where ci = {Tk | Tk ∈ DB} that maximizes intercluster distance and minimizes intracluster variance. More formally ∀ i1, i2, j such that Ti1, Ti2 ∈ ci and Tj ∈ c j D(Ti1, Tj) D(Ti1, Ti2).
There have been numerous approaches for whole series clustering. Typically, after defining an adequate distance function, it is possible to adapt any algorithm provided by the generic clustering topic.

2.2.2. Sub-sequence Clustering. In this approach, the clusters are created by extracting sub-sequences from a single or multiple longer time series.

Definition 2.2.2. Given a time series T = (t1, . . . , tn) and a similarity measure D(Q, C), find the set of clusters C = {ci} where ci = {Tj | Tj ∈ SnT } is a set of sub-sequences that maximizes inter-cluster distance and intracluster cohesion.

2.3 classification

The classification task seeks to assign labels to each series of a set. The main difference when compared to the clustering task is that classes are known in advance and the algorithm is trained on an example dataset. The goal is first to learn what are the distinctive features distinguishing classes from each other. Then, when an unlabeled dataset is entered into the system, it can automatically determine to which class each series belongs. Classification maps input data into predefined groups. It is often referred to as supervised learning, as the classes are determined prior to examining the data; a set of predefined data is used in training process and learn to recognize patterns of interest. Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes.

Two most popular methods in time series classification include the Nearest Neighbor classifier and Decision trees.

Nearest Neighbor method applies the similarity measures to the object to be classified to determine its best classification based on the existing data that has already been classified. For decision tree, a set of rules are inferred from the training data, and this set of rules is then applied to any new data to be classified. Note that even though decision trees are defined for real data, attempting to apply raw time series data could be a mistake due to its high dimensionality and noise level that would result in deep, bushy tree.

2.4 Segmentation

The segmentation (or summarization) task aims at creating an accurate approximation of time series, by reducing its dimensionality while retaining its essential features. Segmentation in time series is often referred to as a dimensionality reduction algorithm. Although the segments created could be polynomials of an arbitrary degree, the most common representation of the segments is of linear functions. Intuitively, a

Definition 2.4.1. Given a time series T = (t1, . . . , tn), construct a model T^{-} of reduced dimensionality d^{-} (d^{-} n) such that T^{-} closely approximates T . More formally $|R(T^{-}) - T| < r$, $R(T^{-})$ being the reconstruction function and r an error threshold.

Piece-wise Linear Representation (PLR) refers to the approximation of a time series Q, of length n, with K straight lines.

2.5 Prediction

Time series are usually very long and considered smooth, that is, subsequent values are within predictable ranges of one another. The task of prediction is aimed at explicitly modeling such variable dependencies to forecast the next few values of a series.

Definition 2.5. Given a time series T = (t1, . . . tn), predict the k next values (tn+1, . . . , tn+k) that are most likely to occur.

2.6 Anomaly Detection

In time series data mining and monitoring, the problem of detecting anomalous/surprising/novel patterns has attracted much attention. In contrast to sub-sequence matching, anomaly detection is identification of previously unknown

patterns. The problem is particularly difficult because what constitutes an anomaly can greatly differ depending on the task at hand. In a general sense, an anomalous behavior is one that deviates from "normal" behavior.

Definition 2.6. Given a time series T = (t1, . . . , tn) and a model of its normal behavior, find all sub-sequences T ∈ SnT that contain anomalies, that is, do not fit the model.

2.7 Motif Discovery

Motif discovery consists in finding every sub-sequence (named motif) that appears recurrently in a longer time series. This idea was transferred from gene analysis in bio-informatics.

Definition 2.7. Given a time series T = (t1, . . . , tn), find all subsequences T ∈ SnT that occur repeatedly in the original time series.

3. Time Series Representation

Time series are essentially high-dimensional data. Defining algorithms that work directly on the raw time series would therefore be computationally too expensive. The main motivation of representations is thus to emphasize the essential characteristics of the data in a concise way. Additional benefits gained are efficient storage, speedup of processing, as well as implicit noise removal. These basic properties lead to the following requirements for any representation:

—significant reduction of the data dimensionality;
—emphasis on fundamental shape characteristics on both local and global scales;
—low computational cost for computing the representation;
—good reconstruction quality from the reduced representation;
—insensitivity to noise or implicit noise handling.Suggestion

4. Application of time series Data mining

Below is a list of few possible ways to take advantage of time series datasets:

- **Trend analysis:** Just plotting data against time can generate very powerful insights. One very basic use of time-series data is just understanding temporal pattern/trend in what is being measured.
- **Outlier/anomaly detection:** An outlier in a temporal data-set represents an anomaly. Whether desired (e.g. profit margin) or not (e.g. cost), outliers detected in a dataset can help prevent unintended consequences.
- **Forecasting:** Forecasting future values using historical data is a common methodological approach – from simple extrapolation to sophisticated stochastic methods such as ARIMA.
- **Predictive analytics:** Advanced statistical analysis such as panel data models (fixed and random effects models) rely heavily on multi-variate longitudinal datasets. These types of analysis help in business forecasts, identify explanatory variables, or simply help understand associations between features in a dataset.