# Time Series Data Mining Tool

Samir Sheriff, Satvik N, Vaishakh BN

## Introduction

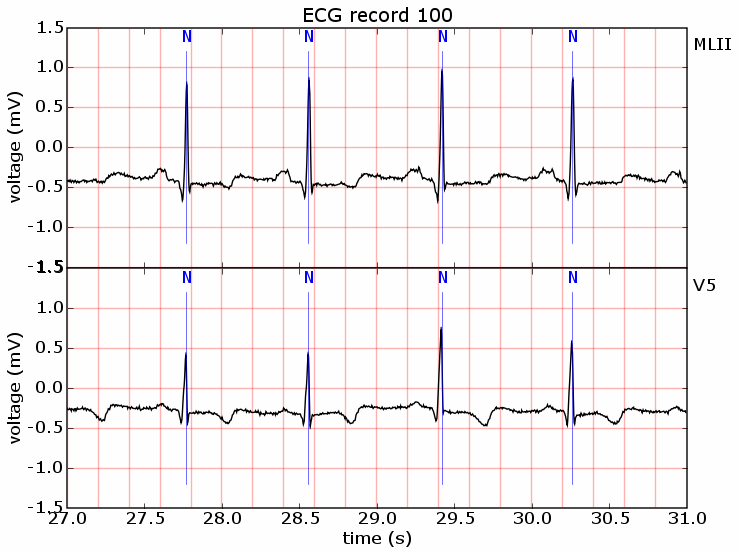
A **time series** is a set of observations *Xt* , each one being recorded at a specific time *t* . A *discrete-time time series*  is one in which the set *T* of times at which observations are made is a discrete set. *Continuous-time time series* are obtained when observations are recorded continuously over some time interval, e.g., when *T*0 belongs [0*,*1]. Examples of time series are the daily closing value of the ECG readings and the annual flow volume of the Nile River at Aswan. Time series are very frequently plotted via line charts.

**Time series analysis** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series forecasting** is the use of a model to predict future values based on previously observed values.

Time series analysis can be applied to:

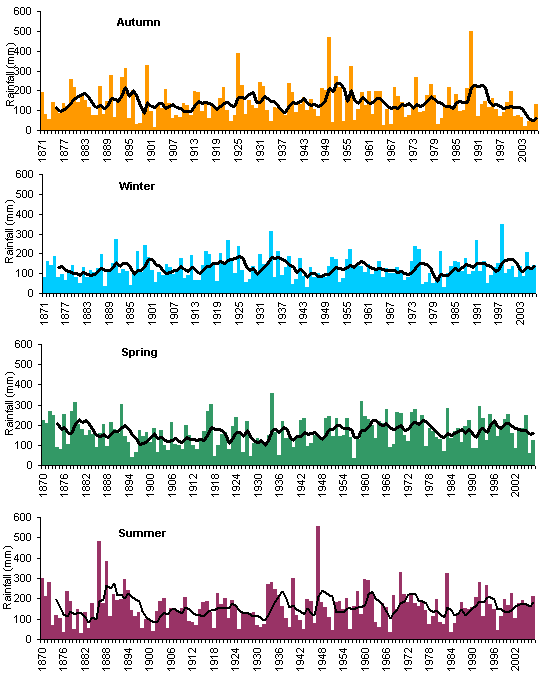
* real-valued, continuous data
* discrete numeric data
* discrete symbolic data (i.e. sequences of characters, such as letters and words in English language).

### Examples

A time series is a set of statistics, usually collected at regular intervals. 

Time series data occur naturally in many application areas.

* Economics - e.g., monthly data for unemployment, hospital admissions, etc.
* Finance - e.g., daily exchange rate, a share price, etc.
* Environmental - e.g., daily rainfall, air quality readings.
* Medicine - e.g., ECG brain wave activity every 2−8 secs.   
  Below are some of the examples of time series data.



## Objectives

The examples considered in the previous section are an extremely small sample from the multitude of time series encountered in the fields of engineering, science, sociology, and economics. Our objectives in this project is to study techniques for drawing inferences from such series. It is necessary to set up a hypothetical probability model to represent the data. After an appropriate family of models has been chosen, it is then possible to estimate parameters, check for goodness of fit to the data, and possibly to use the fitted model to enhance our understanding of the mechanism generating the series. Once a satisfactory model has been developed, it may be used in a variety of ways depending on the particular field of application.

Below are the major tasks of the time series data mining.

1. Indexing - Given a query time series Q, and some similarity/dissimilarity measure D(Q;C), find the most similar time series in database DB
2. Clustering - Find natural groupings of the time series in database DB under some similarity/dissimilarity measure D(Q;C)
3. Classification - Given an unlabeled time series Q, assign it to one of two or more predefined classes
4. Prediction - Given a time series Q containing n data points, predict the value at time n + 1.
5. Summarization - Given a time series Q containing n data points where n is an extremely large number, create a (possibly graphic) approximation of Q which retains its essential features but fits on a single page, computer screen, etc.
6. Anomaly Detection - Given a time series Q, assumed to be normal, and an unannotated time series R, find all sections of R which contain anomalies or “surprising/interesting/unexpected” occurrences.
7. Segmentation - Given a time series Q containing n data points, construct a model Q, from K piecewise segments (K << n), such that Q closely approximates Q.

Time series models are also useful in simulation studies. For example, the performance of a reservoir depends heavily on the random daily inputs of water to the system. If these are modeled as a timeseries, then we can use the fitted model to simulate a large number of independent sequences of daily inputs.

## Literature Survey

A time series is a collection of observations made sequentially through time. At each time point one or more measurements may be monitored corresponding to one or more attributes under consideration. The resulting time series is called univariate or multivariate respectively. In many cases the term sequence is used in order to refer to a time series, although some authors refer to this term only when the corresponding values are non-numerical. Throughout this paper the terms sequence and time series are being used interchangeably.

As mentioned in the previous section, the most common tasks of time series data mining methods are: indexing, clustering, classification, novelty detection, motif discovery and rule discovery. In most of the cases, forecasting is based on the outcomes of the other tasks. A brief description of each task is given below.

Indexing: Find the most similar time series in a database to a given query time series.

Clustering: Find groups of time series in a database such that, time series of the same group are similar to each other whereas time series from different groups are dissimilar to each other.

Classification: Assign a given time series to a predefined group in a way that is more similar to other time series of the same group than it is to time series from other groups.

Novelty detection: Find all sections of a time series that contain a different behavior than the expected with respect to some base model.

Motif discovery: Detect previously unknown repeated patterns in a time series database.

Rule discovery: Infer rules from one or more time series describing the most possible behaviour that they might present at a specific time point (or interval).

The temporal aspect of data arises some special issues to be considered and/or imposes some restrictions in the corresponding applications. First, it is necessary to define a similarity measure between two time series and this issue is very important in TSDM since it involves a degree of subjectivity that might affect the final result. A lot of research has focused on defining different similarity measures in order to improve the performance of the corresponding methods. Second, it is necessary to apply a representation scheme on the time series data. Since the amount of data may range from a few megabytes to terabytes, an appropriate representation of the time series is necessary in order to manipulate and analyze it efficiently. The desirable properties that this approach should hold are:

(a) the completeness of feature extraction

(b) the reduction of the dimensionality “curse” [ 1].

More specifically, the method of extraction features should guarantee that there would be no pattern missed, the number of patterns falsely identified as interesting will be minimized and the dimensionality reduction will be substantial. In many cases also, the objective is to take advantage of the specific characteristics of a representation that make specific methods applicable (i.e. inducing rules, Markov models). Consequently, the majority of the researchers are focused on defining novel similarity measures and representation schemes in order to improve indexing performance.

Clustering and classification of time series rely heavily on the similarity measure and the representation scheme selected, thus, there are very few papers proposing a novel algorithm [2]. A recent survey on clustering time series is provided by Liao [3].

Novelty detection is a very important task in many areas. Several alternative terms for “novelty” have been used, such as, “anomaly”, “interestingness”, “surprising”, “faults” to name a few. Moreover, many problems of finding periodic patterns can be considered as similar problems. The important point here is to provide a clear and concise definition of the corresponding notion. For instance, Keogh et al. [14] describe a pattern as surprising “if the frequency with which it appears, differs greatly from that expected given previous experience”. The authors present a novel algorithm, called Tarzan, and provide useful pointers to relevant literature. Recently, Aref et al. [4] focus on discovering partial periodic patterns in one or more databases. They present algorithms for incremental mining (how to maintain discovered patterns over time as the database is being expanded).

Motif discovery has only recently attracted the interest of the data mining community [9]. Motifs are defined to be previously unknown, frequently occurring patterns in a time series. These patterns may be of particular importance to other data mining tasks, such as, rule discovery and novelty detection. The recent work of Tanaka et al. [7] proposes a new method for identifying motifs from multi-dimensional time series. They apply Principal Component Analysis to reduce dimensionality and perform a symbolic representation. Then, the motif discovery procedure starts by calculating a description length of a pattern based on the “Minimum Description Length” principle.

### 1. Indexing

Indexing approaches are mostly influenced by the pioneer work of Agrawal et al. [1], generalized by Faloutsos et al. [12]. The emerged framework from these papers, referred as GEMINI, can be summarized in the following steps [11]:

· extract k essential features from the time series

· map into a point in k-dimension feature space

· organize points with off-the-shelf spatial access method (‘SAM’)

· discard false alarms

The first and second step suggests the application of a representation scheme in order to reduce the dimensionality. However, this mapping should guarantee that it would return all the qualifying objects. This implies that the similarity measure in the k-dimension feature space should lower bound the corresponding similarity measure in the original space [8]. The third step is an opened selection, however most of the times R-tree structures are used. Other indexing structures may be vp-trees [7] [9], hB-trees and grid-files. The fourth step is a consequence of the fact that this approach can not guarantee that there will not be returned unqualified objects, thus these false alarms should be discarded in a post processing phase.

Recently, Vlachos et al. [8] presented an external memory indexing method for discovering similar multidimensional time series under time warping conditions. The main contribution of this work is the ability to support various distance measures without the need to reconstruct the index. Two approaches with respect to distance measures are taken under consideration, namely, the Longest Common Subsequence (LCS) and the Dynamic Time Warping (DTW). Their indexing technique works by splitting a set of multiple time series in multidimensional Minimum Bounding Rectangles (MBR) and storing them in an R-tree. For a given query, a Minimum Bounding Envelope (MBE) is constructed, that covers all the possible matching areas of the query under time warping conditions. This MBE is decomposed into MBRs and then probed in the R-tree index.

### 2. Time series representation

There have been several time series representations proposed in the literature, mainly on the purpose of reducing the intrinsically high dimensionality of time series. We will refer to some of the most commonly used representations. Discrete Fourier Transform (DFT) [1] was one of the first representation schemes proposed within data mining context. DFT transforms a time series from the time domain into the frequency domain whereas a similar representation scheme, Discrete Wavelet Transform (DWT) [8], transforms it into the time/frequency or space/frequency domain. Singular Value Decomposition (SVD) [5] performs a global transformation by rotating the axes of the entire dataset such that the first axis explains the maximum variance, the second axis explains the maximum of the remaining variance and is orthogonal to the first axis etc. Piecewise Aggregate Approximation (PAA) [3] divides a time series into segments of equal length and records the mean of the corresponding values of each one. Adaptive Piecewise Constant Approximation (APCA) [10] is similar to PAA but allows segments of different lengths. Piecewise Linear Approximation (PLA) approximates a time series by a sequence of straight lines.

Recently, more representation schemes have been proposed in order to reduce dimensionality. The first class of these schemes consists of symbolic representations. Lin et al. [11] propose a Symbolic Aggregate Approximation (SAX) method, which uses as a first step the PAA representation and then discretizes the transformed time series by using the properties of the normal probability distribution. Bagnal[5] assess the effects of clipping original data on the clustering of time series. Each point of a series is mapped to 1 when it is above the population mean and to 0 when it is below. This representation is called clipping and has many advantages especially when the original series is long enough. It achieves adequate accuracy in clustering, it efficiently handles outliers and it provides the ability to employ algorithms developed for discrete or categorical data. Megalooikonomou et al. [30] introduce a novel dimensionality reduction technique, called Piecewise Vector Quantized Approximation (PVQA). This technique is based on vector quantization that partitions each series into segments of equal length and uses vector quantization to represent each segment by the closest codeword from a codebook. The original time series is transformed to a lower dimensionality series of symbols. This approach requires a training phase in order to construct the codebook, a data-encoding scheme and a distance measure.

Cole et al. [9] provide a work that addresses the task of discovering correlated windows of time series (synchronously or with lags) over streaming data. They concentrate in the case where the time series are “uncooperative”, meaning that there does not exist a fundamental degree of regularity that would allow an efficient implementation of DFT transformations. The proposed method involves a combination of several techniques – sketches (random projections), convolution, structured random vectors, grid structures, and combinatorial design – in order to achieve high performance. Gionis and Mannila [7] introduce a different approach, which is mainly motivated from research on human genome sequences. However, this approach is more general and involves multivariate time series. The notion behind their approach is that, the high variability that some time series very often exhibit, may be explained by the existence of several different sources that affect different segments of this series. More specifically, the task is to find a proper way to segment a time series into k segments, each of which comes from one of h different sources (k >>h). This task is analogous to clustering the points of a time series in h clusters with the additional constraint that a cluster may change at most k-1 times. Gionis and Mannila provide three algorithms for solving this problem and they test them on synthetic and genome data.

Finally, Vlachos et al. [3] propose to represent a time series by applying discrete Fourier transformations and retain the k best Fourier coefficients instead of the first few ones. Although this paper is motivated by mining knowledge from the query logs of the MSN search engine, the proposed methods may be applied for time series data mining in general.

### 3. Similarity Measures

The definition of novel similarity measures has been one of the most researched areas in the TSDM field. Generally, they are strongly related to the representation scheme applied to the original data. However, there are some similarity measures that appear frequently in the literature. Most of the researchers’ choices are based on the family of Lp norms, that include the Euclidean distance. Yi and Faloutsos [3] presented a novel and fast indexing scheme when the distance function is any of the arbitrary Lp norms (p = 1, 2, …, ¥). Another similarity measure that attracted a lot of attention, Dynamic Time Warping (DTW), comes from the speech recognition field [6]. The main advantage of this measure is that it allows acceleration-deceleration of a series along the time dimension (nonlinear alignments are possible), however it is computationally expensive. Markov models have been constructed and experimented. Another family of distance measures, **Longest Common Subsequence Measures** (LCS), often used in speech recognition and text pattern matching. As an example of this approach, we refer to the work of Agrawal et al. [2] who define two sequences as similar when they have enough, non-overlapping, time-ordered pairs of subsequences that are similar.

Li et al. [6] propose an algorithm for fast and efficient recognition of motions in multi-attribute continuous motion sequences. The main contribution of this paper is the definition of a similarity measure based on the analysis of Singular Value Decomposition (SVD) properties of similar multi-attribute motions. The proposed measure deals with noise and takes into account the different rates and durations of each motion. The authors also propose a five-phase algorithm for handling segmentation and recognition in real-time.

Sakurai et al. [5] propose the Fast search method for dynamic Time Warping (DTW) that satisfies the following criteria:

(a) it is fast

(b) it produces no false dismissals

(c) it does not pose any restriction on the series length

(d) it supports for any, as well as for no restriction on warping scope.

Their approach is based on a new lower bounding distance measure. They represent the sequence with approximate segments, not necessary of equal length, and operate on them. Three segments, the lower bound, the upper bound, and the time interval, correspond to each one of these approximate segments. In order to fulfill all of the above criteria, they provide algorithms for dynamic programming and searching adjusted to the properties of this representation. Fu et al. [14] propose a new technique to query time series that incorporates global scaling and time warping. The argument is that most real world problems require the ability to handle both types of distortion simultaneously. The approach is to scale the sequence by a bounded scaling factor and also to find nearest neighbor or evaluate range query by applying time warping. The authors provide definitions and proofs of the necessary lower bounds.

Furthermore, there is the expected contribution to defining similarity measures by papers that propose novel representation schemes, since these two tasks are interrelated to each other.

References :

[1] Agrawal R., Faloutsos C., Swami A., “Efficient Similarity Search In Sequence Databases”, Proc.

FODO 1993, pp.69-84, Chicago, Illinois (USA), October 1993.

[2] Agrawal R., Lin K.-I., Sawhney H. S., Shim K., “Fast Similarity Search in the Presence of Noise,

Scaling, and Translation in Time-Series Databases”, Proc. VLDB 1995, pp. 490-501, Zurich

(Switzerland), September 1995.

[3] Antunes C.M., Oliveira A.L., “Temporal Data Mining: An Overview”, Workshop on Temporal Data

Mining ACM SIGKDD 2001, San Francisco, California (USA), August 2001.

[4] Aref W. G., Elfeky M. G., Elmagarmid A. K., “Incremental, Online, and Merge Mining of Partial

Periodic Patterns in Time Series Databases”, IEEE Transactions on Knowledge and Data

Engineering, vol. 16, no. 1, pp. 332-342, January 2004.

[5] Bagnall A. J., Janacek G. J., “Clustering Time Series from ARMA Models with Clipped Data”, Proc.

ACM SIGKDD 2004, pp. 49-58, Seattle, WA (USA), August 2004.

[6] Berndt D. J., Clifford J., “Using Dynamic Time Warping to Find Patterns in Time Series”, KDD

Workshop 1994, pp.359-370, Seattle, WA (USA), July 2004.

[7] Bozkaya T., Yazdani N., Ozsoyoglu M., “Matching and Indexing Sequences of Different Lengths”,

Proc. CIKM 1997, pp. 128-135, Las Vegas, Nevada (USA), November 1997.

[8] Chan K., Fu A. W., “Efficient Time Series Matching by Wavelets”, Proc. ICDE 1999, pp. 126-133,

Sydney (Australia), March 1999.

[9] Cole R., Shasha D., Zhao X., “Fast Window Correlations Over Uncooperative Time Series”, Proc.

ACM SIGKDD 2005, pp. 743-749, Chicago, Illinois (USA), August 2005.

[10]Das G., Lin K.I., Mannila H., Ranganathan G., Smyth P., “Rule Discovery from Time series”, Proc.

ACM SIGKDD 1998, pp. 16-22, New York, New York (USA), August 1998.

[11] Faloutsos C., “Mining Time Series Data”, Tutorial ICML 2003, Washington DC (USA), August 2003.

[12]Faloutsos C., Ranganathan M., Manolopoulos Y., “Fast Subsequence Matching in Time Series

Databases”, Proc. ACM SIGMOD 1994, pp. 419-429, Minneapolis, Minnesota (USA), May 1994.

[13]Fayyad U.M., Piatetsky-Shapiro G., Smyth P., “From Data Mining to Knowledge Discovery: An

Overview”, In Advances in knowledge Discovery and Data Mining, pp. 1-30, AAAI / MIT Press 1996.

[14] Eamonn Keogh, Shruti Kasetty., “On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration

|  |  |
| --- | --- |
|  |  |
|  | |