# Current Trends in Time Series Representation

Leonidas Karamitopoulos, Georgios Evangelidis

Dept. of Applied Informatics, University of Macedonia, Thessaloniki, Greece
Tel: +302310-891844, Email: {lkaramit, gevan}@uom.gr

## Abstract

Time series data generation has been exploded in almost every domain such as in business, industry, medicine, science or entertainment. Consequently, there is an increasing need for analysing efficiently the huge amount of this information either online or offline. The inherent characteristics of time series data, specifically, the high dimensionality, the high feature correlation and the large amounts of noise led researchers to focus mostly on proposing novel representation schemes for this type of data in order to address the resulting problems. In addition to that, representation schemes have also been proposed for the purpose of exploiting specific methods and algorithms that require different type of data, such as Markov models. Concurrently, since most of the data mining tasks require searching for similar patterns, such as query by content, clustering or classification, many papers have been published that propose a generic or representation-based similarity measure. The objective of this paper is to serve as an overview of the recent approaches and trends in the field of time series data mining representations. It seems that the main interest has been gradually drawn in manipulating streaming data and / or multivariate time series. Although a general overview is included, the literature review is focused mainly on papers of the last four years.

**Keywords:** time series, representations, data mining

## *1. Introduction*

The Data Mining (DM) field has attracted a lot of attention during the last decade since it involves techniques and algorithms capable of efficiently extracting patterns that can potentially constitute knowledge from large databases. Time Series Data Mining (TSDM) is a relatively new field that comprises DM methods adjusted in a way that they take into consideration the temporal nature of data. Several procedures generate huge amounts of data in the form of time series in almost every domain, for example, business, industry, medicine and science. In addition to that, TSDM techniques can be applied on image or video data since these types of data can be considered also as time series. According to the research in this field, the main tasks of TSDM methods are: indexing, clustering, classification, novelty detection, motif discovery and rule discovery. Although some of these tasks are similar to the corresponding DM tasks, the temporal aspect arises two special issues to be considered, namely, the representation of the time series and the definition of a similarity measure between two time series.

To our knowledge, there is no paper reviewing thoroughly the Time Series Data Mining field. However, there exist some excellent tutorials (i.e. [Keogh (2003)]) and also, a presentation of results from extensive experiments on a wealth of TSDM techniques [Keogh & Kasetty (2002)]. The objective of our paper is to serve as an overview of the most recent time series representations, which constitute a basic component of all TSDM methods. The survey cannot be considered comprehensive since the emerging field of data mining has attracted the interest of researchers from many diverse fields such as computing science, bioinformatics, manufacturing etc. Nevertheless, the reviewed papers come from high quality conferences and journals, including SIGKDD, SIGMOD, VLDB, and CIKM, from the past four years.

In Section 2 we briefly provide the background of the Time Series Data Mining field. In Section 3, we present the various types of time series representations proposed to date. Finally, a conclusion is presented in Section 4.

## *2. Time Series Data Mining Background*

A time series is a collection of real-valued observations made sequentially through time. At each time point one or more measurements that correspond to one or more attributes under consideration are monitored. The resulting time series are referred as univariate (one-dimensional) or multivariate (multi-dimensional) respectively.

### *2.1 Time Series Data Mining Tasks*

The most common tasks of TSDM methods are: indexing, clustering, classification, novelty detection, motif discovery and rule discovery. **Indexing** is a core task to almost all TSDM methods and refers to efficiently finding the most similar time series in a database to a given query time series. **Classification** involves assigning a given time series to a predefined group of similar time series. **Clustering** groups similar time series when predefined groups are not available. **Novelty** (or anomaly) **detection** finds all sections of a time series that exhibit a different behavior than the expected with respect to some base model. Many problems of finding periodic patterns can be considered as similar problems. **Motif discovery** refers to detecting previously unknown repeated patterns in a time series database. Finally, **rule discovery** infers rules from one or more time series describing their most possible behaviour at a specific time point (or interval).

### *2.2 Time Series Data Mining Issues*

The temporal nature of time series arises some special issues to be considered and/or imposes some restrictions in the corresponding applications. First, it is necessary to transform the original time series data into a feature space by applying a representation scheme for two reasons: (a) to make specific methods applicable (i.e. rule induction, Markov models) or (b) to reduce dimensionality. The latter reason is the most important one in the data mining context, since it is necessary to manipulate and analyze efficiently huge amount of data that may range from a few megabytes to terabytes. The desirable properties in this case are: (a) the completeness of feature extraction and (b) the reduction of the dimensionality "curse" [Agrawal et. al. (1993]. More specifically, the method of extracting the essential features of the original time series should guarantee that there would be no pattern missed and the number of patterns falsely identified as interesting will be minimized.

Second, since most of the data mining tasks require the identification of similar time series, it is necessary to define a similarity measure. This issue interrelates to the previous one in that the selected similarity measure and representation should guarantee the completeness of feature extraction. In [Faloutsos et. al. (1994)] the so-called *bounding lemma* is presented, which intuitively states that in order to guarantee the completeness of feature extraction, the distance between two time series in the feature space should match or underestimate their true distance. The definition of novel similarity measures has been one of the most researched areas in the TSDM field. Most of the researchers' choices, however, are based on the family of Lp norms that include the Euclidean distance [Yi & Faloutsos (2000)] and the Dynamic Time Warping (DTW) [Berndt & Clifford (1994)]. A thorough tutorial on similarity measures along with other TSDM related issues is provided in [Gunopoulos & Das (2000)]. Recently, the research on similarity measures is focused, mainly, on improving existing measures such as DTW [Sakurai et al. (2005)], as well as, on defining or adjusting distance measures for multivariate time series and streaming data [Yang & Shahabi (2004)] and [Li et al. (2004)]. Finally, a new technique to query time series that incorporates global scaling and time warping is proposed in [Fu et al. (2005)].

## 3. 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. A hierarchy of various time series representations is presented in a tree diagram in [Lin et. al. (2003)].

Within data mining context, one of the first representation schemes proposed was the Discrete Fourier Transform (DFT) [Agrawal et. al. (1993)], which is based on the transformation of a time series from the time domain into the frequency domain. Another similar approach, Discrete Wavelet Transform (DWT) transforms a time series into the time/frequency [Chan & Fu (1999)]. Singular Value Decomposition (SVD) [Korn et. al. (1997)] performs a global transformation by rotating the axes of the entire dataset and represents the time series as a linear combination of the most important principal components. Piecewise Aggregate Approximation (PAA) [Yi & Faloutsos (2000)] divides a time series into segments of equal length and records the mean of the corresponding values of each one, whereas, Piecewise Linear Approximation (PLA) approximates a time series by a sequence of straight lines [Shatkay & Zdonik (1996)].

*Table 1.* ***Taxonomy of Reviewed Papers***

| **TSDM Tasks** | **Univariate** | **Multivariate** |
|---|---|---|
| **General** | [Lin (2003)] (**str**)<br>[Fu (2004)] (**str**)<br>[Vlachos (2004a)]<br>[Megalooikonomou (2005)]<br>[Toshniwal (2005)] | [Gionis (2003)]<br>[Cole (2005)] (**str**) |
| **Clustering / Classification** | [Bagnal & Janacek (2004)]<br>[Kontaki (2005)] (**str**) | [Jenkins (2003)]<br>[Vlachos (2004b)] |
| **Novelty Detection** | [Morchen (2005)] | [Sayal (2004)] (**str**)<br>[Fujimaki (2005)]<br>[Papadimitriou (2005)] (**str**) |
| **Motif Discovery** | [Vlachos (2004a)] | [Tanaka (2005)] |
| **Rule Discovery** | [Morchen (2005)] | [Sayal (2004)] (**str**) |

Recently, new representation schemes have been proposed in order to reduce dimensionality. It seems that the increasing interest in analyzing streaming data and / or multivariate time series led researchers to propose novel approaches to representation or to adjust previously proposed ones. Moreover, several of these approaches are driven by specific TSDM tasks. According to these observations, we present a taxonomy of the reviewed papers with respect to the corresponding TSDM task and the dimensionality of the time series. This does not mean that the proposed representation in a paper classified under "Univariate" column cannot be extended for multivariate time series or other TSDM tasks. In addition to that, an indicator (str) appears next to each paper indicating whether the procedure of gathering data (streaming or not) is taken into consideration.

### *3.1 Univariate Time Series Representation*

A broad class of time series representations involves symbolic representations. The *Symbolic Aggregate Approximation* (SAX) method is proposed in [Lin et. al. (2003)]. SAX uses as a first step the PAA representation and then discretizes the transformed time series by using the properties of the normal probability distribution. The resulting "word" depends on a previously chosen alphabet size. Another symbolic representation, called *Persist*, is provided in [Morchen & Ultsch (2005)]. This is a new unsupervised discretization method of time series, which results into a sequence with symbols that retain their temporal aspect. This method requires that the time series does not change behavior fast and does not contain a long-term trend. The basic idea of this approach is that a time series is a sequence of states generated by an underlying process. Persist is based on the *Kullback_Leibler divergence* between the marginal and the self-transition probability distributions of the states (the discretization symbols) in order to discover these states. In [Bagnal & Janacek (2004)] the effects of *clipping* original data on the clustering of time series is assessed. 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. The authors evaluate the effects of this representation on clustering. Another novel dimensionality reduction technique, called *Piecewise Vector Quantized Approximation* (PVQA), is introduced in [Megalooikonomou et al. (2005)]. 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 (the *Generalized Lloyd Algorithm* is applied), a data-encoding scheme and a distance measure.

Kontaki et al. [Kontaki et al. (2005)] present a method of continuously classifying streaming time series based on their trends. The first step of this method includes the application of Piecewise Linear Approximation (PLA) on smoothed time series under the sliding window paradigm and results in a sequence of pairs (*t*, *trend*), where *t* denotes the first time point of the corresponding segment and *trend* denotes the current trend of it (UP or DOWN). An incremental computation of this representation is presented in order to support the application on streaming data. The classification task is performed based on the identified trends and a set of a priori known classifiers. Fu et al. [Fu et al. (2004)] propose an application-oriented representation scheme with respect to financial time series. They take advantage of the specific head-and – shoulder pattern of interest within stock analysis domain and identify the *perceptually important points* (PIPs). These critical points of higher importance are identified by the vertical distance between the points of interest to their current pair of adjacent PIPs. According to the proposed representation, the points of the original time series are reordered in a way that a data point identified as perceptually important in an earlier stage is considered being more important than those points identified afterwards and are stored in a specialized binary tree. Authors provide comparative experiments on various methods of updating where new values arrive in a streaming fashion.

A novel representation scheme on the purpose of efficiently identifying similar time series is introduced in [Toshniwal & Ramesh (2005)]. Their technique is based on the assumption that similar time series correspond to *centroids* that are close to each other, after properly pre-processing raw data in order to deal with different shifts and scales. The determination of the *centroids* is based on the calculation of the second moments as they provide weights to the locations of the measured parameter along the time axis.

A variation of the original Discrete Fourier Transform appears in [Vlachos et al. (2004)] that proposes to retain the *k* best Fourier coefficients instead of the first few ones.

### *3.2 Multivariate Time Series Representation*

The first class of representation schemes aims at transforming multivariate time series. In [Vlachos et al. (2004)] the problem of similarity search is introduced on multi-dimensional time series, such as trajectories, under different orientations. They propose a novel method of mapping trajectories into a space that is invariant of translation, scaling and rotation. In addition to that, they define a DTW-based (Dynamic Time Warping) similarity measure on this new space. More specifically, their approach involves transforming the spatial coordinates of a trajectory into a sequence of angle/arc-length pairs and normalizing the transformed series by applying a novel iterative modulo normalization technique, which deals with the possible problem of vertical shifts. The proposed similarity measure allows elastic matches and can be efficiently lower bounded.

In a recent paper [Fujimaki et al. (2005)] present a novel anomaly detection method for spacecrafts, based on Kernel Feature Space and directional distribution. Part of their work is to define an anomaly metric. Although this is an application-oriented method, the fact that it requires little a priori knowledge makes it potentially useful to other applications too.

In [Jenkins & Mataric (2003)] a dimensionality reduction technique is proposed for the purpose of revealing spatio-temporal structures from human motion streams and therefore deriving primitive behaviors (vocabulary). They propose an extension of *ISOMAP*, a non-linear dimensionality reduction technique, which incorporates not only the extraction of spatial structures but also the corresponding temporal dependencies. This is a PCA-based method (Principal Component Analysis) that performs the decomposition on a feature space similarity matrix instead of an input space covariance matrix. The extended technique takes into account the temporal dependencies by adjusting the similarity matrix through Weighted Temporal Neighbours. The combination of this type of dimensionality reduction with clustering is proposed in order to determine a human motion vocabulary.

A recent work [Tanaka et. al. (2005)] proposes a new method for identifying motifs from multi-dimensional time series. It applies 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.

A different approach, which is mainly motivated from research on human genome sequences, is introduced in [Gionis & Mannila (2003)]. 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

The second class of representation schemes aims at identifying patterns among streaming multiple time series. Although each one of the time series is univariate, the fact that they are being generated concurrently results in recording multiple measurements at each time point. Depending on the task, this fact may require treating time series in a multivariate way. *SPIRIT* is introduced in [Papadimitriou et al. (2005)] as a new approach to discovering patterns from streaming multiple time series. This approach identifies correlations and hidden variables among time series by applying Principal Component Analysis, in order to summarize the entire set of streams and provide useful means in efficient forecasting. The *SPIRIT* also satisfies the important requirements of an efficient streaming pattern discovery procedure, that is, it is streaming, it scales linearly with the number of time series, it is adaptive to changes and it is automatic.

In [Cole et. al. (2005)] the task of discovering correlated windows of time series (synchronously or with lags) over streaming data is addressed. The authors concentrate on 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 or DWT 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.

In [Sayal (2004)] a novel method for extracting time-correlations among multiple time-series is introduced. The time-correlations are expressed in the form of correlation rules such as: if A increases more than 5% then B decreases more than 10% on the next day. In order to reduce the search space, the proposed method includes summarization of the original data by aggregation at different time granularities and detection of the change points upon which the desirable comparisons will be based. For the latter task, CUSUM, a well known statistical method, is used in order to convert continuous time series data to discrete data. The resulting representation includes the change points along with a label (UP or DOWN) that indicates the direction of the change and the amount of change.

## *4. Conclusion*

We provided an overview of the current trends in time series representation within data mining context. The rapidly increasing generation of time series data necessitates the development of new techniques and tools to analyze them efficiently and accurately. In the core of these techniques lies the representation of time series in order to take advantage of a wealth of methods, which require different types of data, and to deal with the intrinsic high dimensionality.

There are three key observations. First, there is a trend to adjust existed techniques to data mining tasks performed on multivariate time series. The emerging need of data reduction leads the researchers in adopting well-known methods, such as principal component analysis, in their approaches. Second, the increasing demand of analyzing streaming data also results in modifying existing methods and providing new algorithms. Third, there has been a lot of work in exploiting TSDM techniques from diverse application areas.

The fact that there has been a lot of research on the Time Series Data Mining field from several communities, besides computing science, suggests a closer, interdisciplinary cooperation. Future work will focus on broader literature review, covering advances from these communities, with the hope of bringing closer these research communities.

## *References*

Agrawal R., Faloutsos C., Swami A. (1993), E*fficient similarity search in sequence databases*, in Proc. FODO-1993: Fourth International Conference of Foundations of Data Organization and Algorithms, Chicago, Il, USA.

Bagnall A. J., Janacek G. J. (2004), *Clustering time series from ARMA models with clipped data*, in Proc. ACM SIGKDD-2004: Tenth International conference on Knowledge discovery and Data Mining**,** Seattle, WA, USA.

Berndt D. J., Clifford J. (1994), *Using dynamic time warping to find patterns in time series*, KDD-94: AAAI Workshop on Knowledge Discovery in Databases, Seattle, WA, USA.

Chan K., Fu A. W. (1999), E*fficient time series matching by wavelets*, in Proc. ICDE-1999: Fifteenth International Conference on Data Engineering, Sydney, Australia.

Cole R., Shasha D., Zhao X. (2005), *Fast window correlations over uncooperative time series*, in Proc. ACM SIGKDD-2005: Eleventh International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA.

Faloutsos C., Ranganathan M., Manolopoulos Y. (1994), *Fast subsequence matching in time series databases,* in Proc. ACM SIGMOD-1994: International Conference on Management of Data, Minneapolis, MN, USA.

Fu A. W, Keogh E, Lau L. Y. H., Ratanamahatana C. (2005), S*caling and time warping in time series querying*, in Proc. VLDB-2005: Thirty-First International Conference on Very Large Databases, Trondheim, Norway.

Fu T. C., Chung F.L., Luk R. and Ng C.M. (2004), *A specialized binary tree for financial time series representation*, KDD/TDM-2004: Third SIGKDD workshop on mining temporal and sequential data**,** Seattle, WA, USA.

Fujimaki R., Yairi T., Machida K. (2005), A*n approach to spacecraft anomaly detection problem using kernel feature space*, in Proc. PAKDD-2005: Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Hanoi, Vietnam.

Gionis A., Mannila H. (2003), *Finding recurrent sources in sequences,* in Proc. RECOMB-2003: Seventh Annual International Conference on Computational Biology, Berlin, Germany.

Gunopoulos D., Das G. (2000), T*ime series similarity measures*, Tutorial notes ACM SIGKDD-2000: Sixth International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA.

Jenkins O. C., Mataric M. J. (2003), *Automated derivation of behavior vocabularies for autonomous humanoid motion,* in Proc. AAMAS-2003: Second International Joint Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia.

Keogh E., Kasetty S. (2002), O*n the need for time series data mining benchmarks: a survey and empirical demonstration*, in Proc. ACM SIGKDD-2002: Eighth International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada.

Keogh E. (2003), *Data mining and machine learning in time series databases*, Tutorial ECML/PKDD-2003: Fourteenth European Conference on Machine Learning and Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia.

Kontaki M., Papadopoulos A. N., and Manolopoulos Y. (2005), C*ontinuous trend-based classification of streaming time series,* in Proc. ADBIS 2005: Ninth East-European Conference on Advances in Databases and Information Systems, Tallin, Estonia

Korn F., Jagadish H., Faloutsos C. (1997), *Efficiently supporting ad hoc queries in large datasets of time sequences,* in Proc. ACM SIGMOD-1997: International Conference on Management of Data, Tucson, AZ, USA.

Li C., Zhai P., Zheng S. Q., Prabhakaran B. (2004), S*egmentation and recognition of multi-attribute motion sequences,* in Proc. ACM Multimedia-2004: Twelfth Annual Conference, New York, NY, USA.

Lin J., Keogh E., Lonardi S., Chiu B. (2003), *A symbolic representation of time series, with implications for streaming algorithms*, in Proc. DMKD-2003: Eighth Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, CA, USA.

Megalooikonomou V., Li G., Wang Q. (2005), *A dimensionality reduction technique for efficient similarity analysis of time series database*, in Proc. CIKM-2004: Thirteenth International Conference on Information and Knowledge Management, Washington, D.C., USA.

Morchen F., Ultsch A. (2005), *Optimizing time series discretization for knowledge discovery*, in Proc. ACM SIGKDD-2005: Eleventh International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA.

Papadimitriou S., Sun J., Faloutsos C. (2005), *Streaming pattern discovery in multiple time series*, in Proc. VLDB-2005: Thirty-First International Conference on Very Large Databases, Trondheim, Norway.

Sakurai Y., Yoshikawa M., Faloutsos C. (2005), *FTW: Fast search under the time warping distance*, in Proc. PODS: Twenty-Fourth Symposium on Principles of Database Systems, Baltimore, MD, USA.

Sayal M. (2004), *Detecting time correlations in time-series data streams*, Technical Report, Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto HPL-2004-103.

Shatkay H., Zdonik, S.B. (1996), *Approximate Queries and Representations for Large Data Sequences*, in Proc. ICDE: Twelfth International Conference on Data Engineering, New Orleans, LA, USA.

Tanaka Y., Iwamoto K., Uehara K. (2005), *Discovery of time series motif from multi-dimensional data based on mdl principle*, Machine Learning, vol. 58, no. 2-3, pp. 269-300.

Toshniwal D., Ramesh C. J. (2005), *Finding similarity in time series data by method of time weighted moments*, in Proc. ADC-2005: Sixteenth Australasian Database Conference, Newcastle, Australia.

Vlachos M., Meek C., Vagena Z., Gunopoulos D. (2004a), *Identifying similarities, periodicities and bursts for online search queries*, in Proc. ACM SIGMOD-2004: International Conference on Management of Data, Paris, France.

Vlachos M., Gunopulos D. and Das G. (2004b), *Rotation invariant distance measures for trajectories*, in Proc. ACM SIGKDD-2004: Tenth International Conference on Knowledge Discovery and Data Mining**,** Seattle, WA, (USA).

Yang K., Shahabi C. (2004), *A PCA-based similarity measure for multivariate time series*, in Proc. MMDB-2004: Second ACM International Workshop on Multimedia Databases, Arlington, VA, USA.

Yi B. K., Faloutsos C. (2000), *Fast time sequence indexing for arbitrary Lp Norms*, in Proc. VLDB-2000: Twenty-Sixth International Conference on Very Large Databases, Cairo, Egypt.