Journal of Computer Science 10 (12): 2358-2359, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.2358.2359 Published Online 10 (12) 2014 (http://www.thescipub.com/jcs.toc)

Data Mining in Time Series: Current Study and Future Trend

[1]Tan Yan, [2]Liudmila Ulanova, [3]Ye Ouyang and [4]Fengyuan Xu

[1]NEC Laboratories America, Princeton, New Jersey, United States
[2]University of California Riverside, California, United States
[3]Verizon Wireless, Basking Ridge, New Jersey, United States
[4]NEC Laboratories America, Princeton, New Jersey, United States

Time series represent sequences of data points where usually their order is defined by the time when they were recorded. Thus, virtually any sequential recordings can be stored as time series: Stock prices, weather conditions, physical system parameters change, product quality monitoring, etc. This leads to ubiquity of time series in all scientific and practical fields (Esling and Agon, 2012); hence, they have attracted significant research efforts over the past decades. The time series can be univariate, i.e., only one variable recorded, or multivariate, i.e., a set of observations from different sources recorded at some time points. The tasks of time series analysis essentially defined to extract meaningful information from the collections of data points or to organize fast and easy access to the necessary data. In this article, we will briefly discuss the major tasks such as classification, clustering, prediction, segmentation and indexing of time series.

Time series classification is a task that aims to assign each of the time series a class label from two or more classes having some training data. Classification is needed to distinguish between different types of time series. For example, having past recording to separate time series describing yearly weather conditions in different parts of the world or distinguish between types of technical devices where the data were recorded and so forth. The task of classification requires a training set of time series where examples of each class presents. Time series from the testing (unlabeled) set have to be compared with time series in the training set and class labels must be assigned based on some *distance (similarity/dissimilarity) measure* between time series. Distance measures can be based on the shape of time series (e.g., Euclidean distance or Dynamic Time Warping) or on some extracted features (e.g., frequency, amplitude, number of slope changing, etc.). The simple One (or K) Nearest Neighbor (k-NN) rule 0 is highly competitive and used in many domains to assign the class labels for the unlabeled time series. Using the k-NN rule the labels will be assigned according to the labels of the training time series that are the closest (nearest) neighbors to the unlabeled time series in terms of some distance measure. Classification is often referred as *supervised learning* with a special case of the *semi-supervised* learning (Chapelle *et al*., 2006). The latter allows to add the newly-classified time series from the testing set into training set where they can be used for classification of next coming time series. Semi-supervised classification allows mitigating lack of labeled data.

The goal of time series clustering is to group the time series in a dataset according to some principle, or, in other words, to find natural groupings or structure of the dataset. For example, the task may sound as follows: Having some daily recordings of bird sounds from a forest discover the species presented there with no information provided in advance to compare with (Dawson and Efford, 2009). Similarly to the classification task, the time series in the same group (cluster) have to be close according to some distance measure. The clustering may be partitional and hierarchical. Partitional clustering assumes that each cluster has the same level in hierarchy and independent from each other. Hierarchical clustering aims to build a hierarchy of clusters. Clustering is often referred as an *unsupervised learning*, i.e., there are no labeled examples of data that can be used for labels assigning. One of the most popular algorithms of partitional clustering is *k-means clustering*. k-means algorithm starts with choosing *k* centers of clusters (either time series from the dataset or any synthetically generated points) and then, computing the distance to a central time series from the time series in the dataset assigns them to

Corresponding Author: Tan Yan, NEC Laboratories America, Princeton, New Jersey, United States

clusters. The process repeats several times and on each step the central time series is being chosen as a mean value of the time series in the cluster. k-means has some variations and techniques for choice of *k* and number of iterations until convergence. For hierarchical clustering except of having only similarity/dissimilarity measure between instances of time series themselves it is necessary to define a distance measure for similarity of their groups to build a hierarchical structure. Visually hierarchy can be presented as a tree-like structure called dendrogram (Bittmann and Gelbard, 2008).

Time series prediction (forecasting) is used to predict future values of a continuous time series based on the past observations. For example, prediction of stock prices or retail values of goods for the next day taking into account their values in the past weeks. For time series prediction it is necessary to build a model that would describe the behavior of the observed variable over time. The simplest model to employ is a regression model where the values of time series are fitted to a curve with some certain error (Durbin, 1960). The most popular is a linear model fit where the values of time series are fitted to a straight line. However, there are also used higher-level models. If the next value of a variable depends on the previous, as it is usually the case in such fields as, for example, weather and economics, it is suitable to apply *autoregressive* model with integrated moving average. It is usually named as ARIMA (Box and Pierce, 1970).

Segmentation of time series is applied for splitting a time series into pieces with the consequential change of the representation (Keogh *et al*., 2004). This technique is used to alternate the way the time series are represented for faster and easier access. For example, a widely used piecewise approximation divides time series into subparts of equal length and each block of data points is substituted with one value. Another widely used approximation technique SAX (Lin *et al*., 2003) uses the pricewise aggregate approximation of time series as a part of its algorithm. It is often used in indexing of time series ("query by content") that requires finding efficiently similar time series from the dataset to a query object. This approach helps dramatically decrease speed of access to the data.

Concluding our discussion, it is necessary to highlight that time series analysis and data mining is a hot topic in both research community and industrial area. With appearance of lots of different sensors recording time series data that are cheap and easy to use (for instance, any Smartphone is usually equipped with acceleration and gyroscopic sensors.), it becomes easy to obtain time series data. This inevitably leads to explosion of data that require analysis. Therefore, it is likely that high demand for algorithms of big data processing will be even higher in coming decades and the algorithms will have to work on-line, i.e., in real time, to satisfy the needs of the society. Also, having high-performance computers it becomes possible to get more knowledge from the time series with the decrease of computational cost. However, the basic tasks of time series analysis most likely will remain of current importance, but the approaches and techniques must be changed to adjust existing algorithms for big scale data analysis.

References

Bittmann, R.M. and R.M. Gelbard, 2008. DSS Using Visualization of Multi-Algorithms Voting. In: Encyclopedia of Decision Making and Decision Support Technologies, Adam F. and P. Humphreys (Eds.), pp: 297-297.

Box, G.E.P. and D.A. Pierce, 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J. Am. Stat. Assoc., 65: 1509-1526. DOI: 10.1080/01621459.1970.10481180

Chapelle, O., B. Schölkopf and A. Zien, 2006. Semi-Supervised Learning. 1st Edn., MIT Press, Cambridge.

Dawson, D.K. and M.G. Efford, 2009. Bird population density estimated from acoustic signals. J. Appl. Ecol., 46: 1201-1209. DOI: 10.1111/j.1365-2664.2009.01731.x

Durbin, J., 1960. Estimation of parameters in time-series regression models. J. Royal Stat. Soc., 22: 139-153.

Esling, P. and C. Agon, 2012. Time-series data mining. J. ACM Comput. Surveys. DOI: 10.1145/2379776.2379788

Keogh, E.J. S. Chu, D. Hart and M. Pazzani, 2004. Segmenting time series: A survey and novel approach. Data Min. Time Series Databases, 57: 1-22.

Lin, J., E. Keogh, S. Lonardi and B. Chiu, 2003. A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Jun. 13-13, ACM, San Diego, pp: 2-11. DOI: 10.1145/882082.882086