Proc. 13th Int. Conf. on Database Systems for Advanced Applications (DASFAA'08), New Delhi, India, 2008.

Analysis of Time Series Using Compact Model-Based Descriptions

Hans-Peter Kriegel, Peer Kröger, Alexey Pryakhin, and Matthias Renz

Institute for Computer Science
Ludwig-Maximilians-University of Munich
{kriegel,kroegerp,pryakhin,renz}@dbs.ifi.lmu.de

Abstract. Recently, we have proposed a novel method for the compression of time series based on mathematical models that explore dependencies between different time series. This representation models each time series by a combination of a set of specific reference time series. The cost of this representation depend only on the number of reference time series rather than on the length of the time series. In this demonstration, we present a Java toolkit which is able to perform several data mining tasks based on this novel time series representation. In particular, this framework allows the user to explore the properties of our novel approach in comparison to other state-of-the-art compression methods. The results are visually presented in a very concise way so that the user can easily identify important settings of the model-based time series representation.

1 Background

Clustering time series data is a very important data mining task for a wide variety of applications. In many scenarios, the set of time series that need to be analyzed is very large and each time series of this set has an enormous length. Both aspects have obviously a negative influence on the runtime of the clustering process. As a consequence, in many applications, the user is happy with an approximative clustering result, as far as the approximation is not too coarse compared to the original result (effectivity) and the approximative clustering is considerably faster (efficiency) tentatively enabling an interactive analysis. Beside standard compression techniques for similarity search (e.g. [1]), an approach for fast approximative clustering of time series is presented in [2].

Recently, we proposed a novel method for the compression of time series based on mathematical models that explore dependencies between different time series [3]. The resulting representation consists of some low-dimensional feature vector that can easily be indexed by means of any Euclidean index structure. The similarity distance used for the clustering is computed by applying the parameters that specify the combination. Consequently, — in contrast to other existing methods — the cost of the clustering process depend only on the number of reference time series rather than on the length of the time series.

Figure 1 illustrates our approach. A set of reference time series (marked as “T1”, “T2”, and “T3” on the left hand side of Figure 1) is used to approximate an

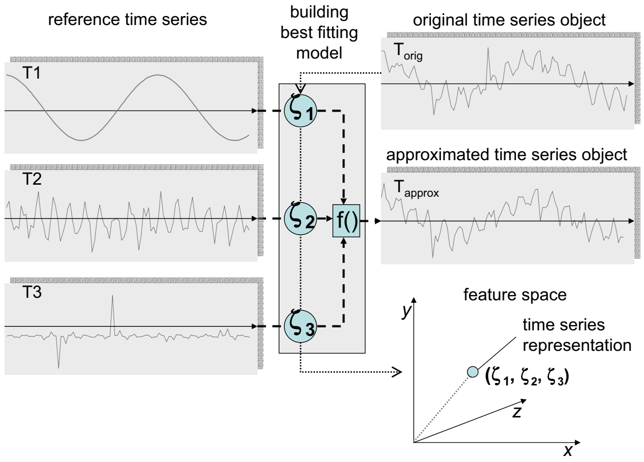


Fig. 1. Model-based time series representation.

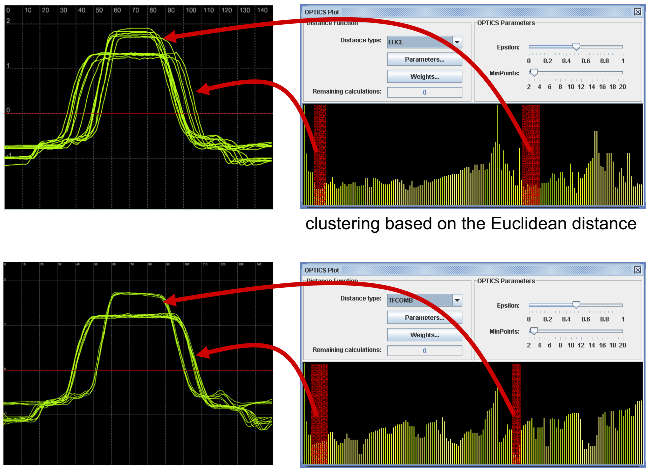
original time series T_{orig}(shown on the upper right hand side of Figure 1) by an arbitrary complex combination T_{approx}. In case of Figure 1 this is a combination of the coefficients $\alpha_1, \ldots, \alpha_3$ representing the three input time series using a function f. The resulting approximated time series (marked as “output” in the middle of the right hand side of Figure 1) is similar to the original time series. For clustering, the approximation is represented by a feature vector of the coefficients of the combination (cf. lower right hand side of Figure 1). Obviously, the choice of a suitable set of reference time series is crucial.

2 A Visual Data Mining Framework for Time Series

The application presented in this work implements a visual data mining approach that analyzes and describes the data in a clear and user-friendly way in order to enable interactive data exploration, in particular cluster analysis. Basically, our framework is designed for the following purposes:

1. It allows visually inspecting the results of comparative data mining, i.e. the results of different compression methods can be evaluated interactively.
2. It assists the user in analyzing the properties of different compression methods, in particular of our model-based technique proposed in [3]. Thus, e.g. the impact of the choice of the reference time series can be evaluated interactively.
3. It is a key tool for a novel step-wise procedure for identifying potentially interesting reference time series for our model-based compression approach.

The concept of our application supports the extraction of novel insights in supervised as well as in unsupervised settings. If class labels are available, the



Distance type: EUCL

Remaining calculations: 0

clustering based on the Euclidean distance

Distance type: TFCOMB

Remaining calculations: 0

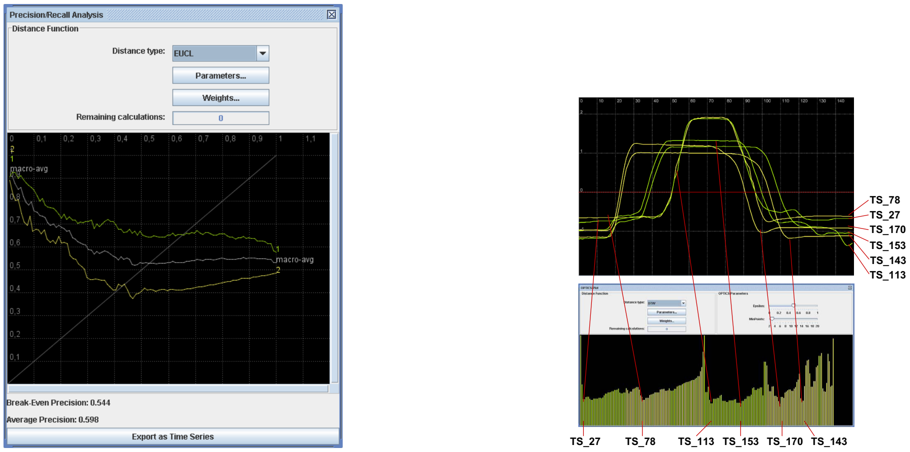
clustering based on mathematical model

Fig. 2. Comparative time series mining.

user can evaluate different settings of reference time series based on classification accuracies in cross-validation experiments. But even in an unsupervised situation where no pre-classified time series are available, our framework can be very helpful. By a quick visual inspection of several clustering results derived for example by OPTICS [4] it is possible to discover important and interesting reference time series based on their ability to form distinct cluster structures.

Comparative Data Mining. In our framework we included several time series similarity measures such as DTW and the Euclidean distance for the purpose of comparing different similarity notions. The collection of implemented distance functions can easily be extended when need arises. Furthermore, most of the existing compression techniques for time series have been implemented including dimensionality reduction techniques as well as our novel model-based compression approach. Our framework enables to visually inspect and comparatively evaluate several clusterings that may be generated using different distance measures or compression models. An illustrative example is visualized in Figure 2 where we compared a clustering based on the Euclidean distance using no compression with the clustering based on our new mathematical model (model-based compression).

Evaluation of Model-based Time Series Compression. The visual clustering output can be used to understand the properties of our novel model-based compression method. For example, it assists in evaluating the impact of the choice of reference time series on the quality of the compression (and, therefore, on the quality of the clustering results) in an interactive way. In a supervised scenario, the user can inspect precision and recall values of different clusterings obtained using different reference time series in order to evaluate the quality of the re-



Distance type: EUCL

Remaining calculations: 0

Break-Even Precision: 0.544

Average Precision: 0.598

(a) Supervised scenario

(b) Unsupervised scenario

Fig. 3. Evaluation of model-based time series compression.

spective results. In an unsupervised scenario, the user can inspect the cluster hierarchies obtained using different reference time series in order to evaluate the quality of the respective results. Screenshots of our framework for both scenarios are visualized in Figure 3.

Determination of Reference Time Series for Model-based Compression. As discussed above, the choice of reference time series is crucial for the quality of data compression and, thus, for the quality of the approximative clustering result. Here, we propose a step-wise procedure for interactively optimizing the set of reference time series based on our novel framework. Starting from an initial clustering of the time series, we can iteratively refine the clustering and the set of reference time series by inspecting the clustering results and updating the set of reference time series accordingly.

References

1. Faloutsos, C., Ranganathan, M., Maolopoulos, Y.: "Fast Subsequence Matching in Time-series Databases". In: Proc. SIGMOD Conference. (1994)
2. Ratanamahatana, C.A., Keogh, E., Bagnall, A.J., Lonardi, S.: "A Novel Bit Level Time Series Representation with Implication for Similarity Search and Clustering". In: Proc. PAKDD Conference. (2005)
3. Kriegel, H.P., Kröger, P., Pryakhin, A., Renz, M., Zherdin, A.: "Approximate Clustering of Time Series Using Compact Model-based Descriptions". In: Proc. DASFAA Conference (to appear). (2008)
4. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: "OPTICS: Ordering Points to Identify the Clustering Structure". In: Proc. SIGMOD Conference. (1999)