Technical Perspective - k-Shape: Efficient and Accurate Clustering of Time Series

Zachary G. Ives
University of Pennsylvania
zives@cis.upenn.edu

Database research frequently cuts across many layers of abstraction (from formal foundations to algorithms to languages to systems) and the software stack (from data storage and distribution to runtime systems and query optimizers). It does this in a way that is specialized to a particular class of data and workloads. Over the decades, we have seen this pattern applied to enterprise data, persistent objects, Web data, sensor data, data streams, and so on. Each time, the community has developed extensions to algebraic query primitives, specialized implementation techniques (index structures, pattern detection algorithms, update and consistency mechanisms, etc.), benchmarks, and new optimization techniques.

Today, we are on the cusp of another class of data and applications becoming of broad interest. *Time series data* were once viewed as being the purview of specialized scientific applications, forecasting settings, etc., with solutions that did not necessarily generalize to other domains. Today time series data are omnipresent: they are continuously emitted by smartphones (tracking location, acceleration, etc.), smartwatches and fitness bands (tracking activity and health data), as well as environmental sensors, medical devices, and flow monitors, and even server logs, network events, financial transactions, etc. Personalization, prediction, and event detection are increasingly reliant on these data.

At first glance, time series data management seems closely related to that of sensor networks, data streams, temporal data, complex event processing, and the like. The difference is largely in the goals: rather than computing properties of samples within time windows, or looking for particular sequences of events, time series processing often looks for general patterns that can manifest themselves with differences in both amplitude and phase: i.e., in some distinguishing *shape* in the readings, which may occur at different scales. For instance, the goal may be to find spikes in voltage levels measured in regions of the brain (EEG, ECoG, etc.) indicative of seizures or tremors, or to find motion in a wrist-worn fitness band that indicates the wearer is taking a step during walking, or to spot a signature in network traffic behavior indicative of a particular kind of DoS attack.

Database and data mining researchers have been developing techniques to extract motifs for time series, which are useful for compression, indexing, and search; query-by-example capabilities with waveforms of particular shapes; techniques to learn the structure of "notable" time series segments (e.g., seizures, tremors); and finally, unsupervised pattern detection methods like clustering, which can be used to find the underlying structure in the shape of the data.

Most work on clustering time series is focused on defining an effective distance metric between shapes, then using standard clustering methods (hierarchical, k-means, k-mediods, spectral) to group time series segments using these distance measures. However, the question of how to describe a shape is not completely evident, and there is also a need to tolerate certain kinds of variations (e.g., noise, stutters, faster or slower rates). As a result, many different measures have been defined, including Euclidean distance functions between waveforms, "time warping" where portions of a time series signal can be accelerated or decelerated, longest common subsequence measures with scaling, and edit distance models where samples can be added or removed to make two shapes look more similar. Excellent survey materials exist describing the different distance measures as well as broader time series techniques [1, 2, 3].

The k-Shape paper is differentiated from previous work by tackling the questions of clustering method and distance measure simultaneously: Paparizos and Gravano use a statistical measure, cross-correlation, as their measure for comparing time series sequences, and they alter the k-means algorithm to use a different centroid computation mechanism when computing the clusters. These two modifications are insightful and highly effective, both in terms of clustering quality and scale-up. In fact, the conference version of the paper includes an extensive performance evaluation with 48 different datasets and shows the superiority of their method against many of the previously proposed schems. This paper represents a very nice example of the progress being made in the time series arena, and we anticipate that many more developments lie ahead.

1. REFERENCES

[1] D. Gunopulos and G. Das. Time series similarity measures and time series indexing. In *SIGMOD*, page 624, 2001.

[2] E. Keogh. Machine learning in time series databases (and everything is a time series!). Tutorial, AAAI 2011. Available from `http://www.cs.ucr.edu/~eamonn/tutorials.html`.

[3] Y. Sakurai, Y. Matsubara, and C. Faloutsos. Mining and forecasting of big time-series data. In *SIGMOD*, pages 919–922, 2015. Tutorial available from `http://www.cs.kumamoto-u.ac.jp/~yasuko/TALKS/15-SIGMOD-tut/`.