Indexing and Mining Streams

Christos Faloutsos
School of Computer Science
CMU
christos@cs.cmu.edu

1. DESCRIPTION - OBJECTIVES

How can we find patterns in a sequence of sensor measurements (eg., a sequence of temperatures, or water-pollutant measurements)? How can we compress it? What are the major tools for forecasting and outlier detection? The objective of this tutorial is to provide a concise and intuitive overview of the most important tools, that can help us find patterns in sensor sequences. Sensor data analysis becomes of increasingly high importance, thanks to the decreasing cost of hardware and the increasing on-sensor processing abilities. We review the state of the art in three related fields: (a) fast similarity search for time sequences, (b) linear forecasting with the traditional AR (autoregressive) and ARIMA methodologies and (c) non-linear forecasting, for chaotic/self-similar time sequences, using lag-plots and fractals. The emphasis of the tutorial is to give the intuition behind these powerful tools, which is usually lost in the technical literature, as well as to give case studies that illustrate their practical use.

NOTICE. : At SIGMOD, Prof. Dennis Shasha will be delivering a related but complementary tutorial, which focuses on anomaly and burst detection in financial and scientific time series.

2. OUTLINE

Similarity Search. We shall cover the need for similarity search; the most popular distance functions (Euclidean, LP norms, time-warping); the most successful indexing methods (R-trees [4], M-trees [2]; and the most popular feature extraction methods from signal processing (DFT, Wavelets, SVD), as well as Multidimensional Scaling and FastMap [3].

Linear Forecasting. We will cover the main idea behind linear forecasting, the popular AR methodology [1] and the related multivariate regression; the powerful Recursive Least Squares [7] method.

Non-linear/chaotic forecasting. We shall present the main idea: the so-called "lag-plots" [6]. Then we will continue with the ubiquitous 'fractals', where we give the definitions of 'fractal dimensions', the intuition behind the major concepts, algorithms for fast computation of fractal dimensions [5]. We conclude with case studies on real datasets.

3. INTENDED AUDIENCE

Researchers that want to get up to speed with the major tools in time sequence analysis. Also, practitioners who want a concise, intuitive overview of the state of the art.

4. REFERENCES

[1] G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1994.

[2] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. *VLDB*, pages 426–435, 1997.

[3] C. Faloutsos. *Searching Multimedia Databases by Content*. Kluwer Academic Inc., 1996.

[4] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. ACM SIGMOD*, pages 47–57, Boston, Mass, June 1984.

[5] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman and Company, New York, 1991.

[6] A. S. Weigend and N. A. Gerschenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison Wesley, 1994.

[7] B.-K. Yi, N. Sidiropoulos, T. Johnson, H. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. *ICDE*, pages 13–22, 2000.

SIGMOD 2004, June 13-18, 2004, Paris, France.