Dimension Reduction Techniques for Identifying Relevant Data Streams

Real-time Analysis of Streaming Data from Sensors

Chandrika Kamath
Lawrence Livermore National Laboratory
7000 East Avenue, Livermore, CA 94551

Abstract

With sensors becoming ubiquitous, there is an increasing interest in mining the data from these sensors as they are being collected. This analysis of streaming data, or data streams, is presenting new challenges for analysis algorithms. The size of the data can be massive, especially when the sensors number in the thousands and the data are sampled at a high frequency. The data can be non-stationary, with statistics that vary over time. The quality of the data may be poor, with noise and missing values. Real time analysis is often required, either to avoid untoward incidents or to understand an interesting phenomena better. These factors make the analysis of streaming data, whether from sensors or other sources, very data- and compute-intensive.

One possible approach to making this analysis tractable is dimension reduction, or the identification of the relevant data streams, so that we can reduce the amount of data that must be monitored. However, what makes a stream relevant often varies from problem to problem. Further, the statistical properties of the data streams can change over time, and a stream identified as important at one instant, may be unimportant at a later instant. As a result, both the number and the set of relevant streams can be variable.

There are two common approaches used to handle the temporal nature of the data and focus on the most recent values - the sliding window approach and the forgetting factor approach. In this presentation, we shall describe ways in which these two ideas have been incorporated into dimension reduction techniques based on correlations, singular value decomposition, and fast subspace trackers. In addition, we shall discuss the results of applying these techniques to one of our test-bed problems, namely, the use of meteorological tower data in the Columbia Basin to predict wind ramp events in the Bonneville Power Administration (BPA) balancing area.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.