Wyatt Blair

EMT-678-WS: Big Data Technologies

10/18/2024

**Class 6 - Stream Data Mining: A Survey**

This paper examines the challenges that come with attempting to data-mind a high-volume data-stream. The data stream has the following three characteristics:

* The data arrives in a continuous stream
* No assumptions on data-stream ordering can be made
* The length of the data is unbounded

There are two main methodologies one could use to make inferences from such a data-stream. One general approach is to take an intelligent subset of the data-stream and make inferences based on that, more manageable, subset. Examples of this methodology include sampling, load-shedding, sketching, and synopsis-data-structures. The other methodology involves making use of, as the author puts it, “algorithms for efficient utilization of space and time”. Examples of this methodology include approximation algorithm, sliding window, and algorithm output granularity.

Once the data is packaged and prepared to make models with, data-scientists have several different techniques they can employ to start making predictions. Techniques like clustering, classification, association, frequency counting, and time series analysis. These techniques and their applications are fairly well understood—typically the hard part of implementing them is preparing a dataset which is clean and organized with which to build these models.