

Data Mining and Environmental Sciences

Shashi Shekhar
University of Minnesota.
shekhar@cs.umn.edu

September 29, 2009

Abstract

It is critical to monitor and predict where and when large contaminant fluxes will occur so that actions may be taken to protect environment and limit exposure to human and aquatic life. Current water quality monitoring is based on infrequent (e.g., weekly) sampling and time-consuming (e.g., hours to days) testing methods, making it difficult to make timely decisions to protect watersheds, a crucial part of our environment. Recent advances have led to use of monitoring networks based on sensors to provide increased sampling frequency as well as digital watershed data warehouses to manage the sensor data. However, key challenges remain. Researchers need new models that they can apply to take full advantage of these new types of data sets, as current models do not adequately account for the huge quantities of data collected and the new patterns that are observable as a result. The goal of this project is to advance new scalable spatio-temporal data mining tools and protocols for monitoring, detecting, and predicting contamination of environment. Classical and spatial data mining ideas are generalized to represent and analyze data sets related to physical processes such as water flow and contaminant flow using novel methods such as flow anomaly detection and teleconnection detection. Given a percentage-threshold and readings from a pair of consecutive upstream and downstream sensors, flow anomaly discovery identifies dominant time intervals where the fraction of time instants of significantly mis-matched sensor readings exceed the given percentage-threshold. Discovering flow anomalies (FA) is an important problem in environmental flow monitoring networks and early warning detection systems for

water quality problems. However, mining FAs is computationally expensive because of the large (potentially infinite) number of time instants of measurement and potentially long delays due to stagnant (e.g. lakes) or slow moving (e.g. wetland) water bodies between consecutive sensors. Traditional outlier detection methods (e.g. t-test) are suited for detecting transient FAs (i.e., time instants of significant mis-matches across consecutive sensors) and cannot detect persistent FAs (i.e., long variable time-windows with a high fraction of time instant transient FAs) due to a lack of a pre-defined window size. In contrast, we propose a Smart Window Enumeration and Evaluation of persistence-Thresholds (SWEET) method to efficiently explore the search space of all possible window lengths. Computation overhead is brought down significantly by restricting the start and end points of a window to coincide with transient FAs, using a smart counter and efficient pruning techniques. Experimental evaluation using a real dataset shows our proposed approach outperforms Naive alternatives.