School of Computer Science
Dissertation
Shiblee Sadik

ONLINE DETECTION OF OUTLIERS FOR DATA STREAMS

ABSTRACT

In applications, such as Web clicks and environmental monitoring, data are in the form of streams, each of which is an infinite sequence of data points with explicit or implicit timestamps and has special characteristics, such as transiency, uncertainty, dynamic data distribution, multi-dimensionality, asynchronous data arrival, dynamic relationships, and schema heterogeneity of data from different sources. In those applications, outliers do exist due to many reasons including human error, instrument error, catastrophe, and malicious behaviour. Being able to detect outliers effectively is critical to many data management and mining tasks. However, not much research has been conducted to discover outliers in data stream applications, especially for those involving multi-dimensionality, related, heterogeneous, and asynchronous streams.

In this dissertation, two innovative outlier detection algorithms, Orion and Wadjet, which take all the data streams' characteristics into consideration are presented. Orion is designed for applications where data are independent streams. It looks for a projected dimension that reveals the outlier nature of multi-dimensional data points with the help of an evolutionary algorithm, and identifies a data point as an outlier if it resides in a low density region in that dimension. Wadjet is designed for applications where data are related, heterogeneous, and asynchronous streams. It has two phases: in the first phase, it processes each data point independently like Orion; and in the second phase, it captures and continuously evaluates the cross-correlation among the data points from multiple streams, and identifies a data point as an outlier if its value does not conform to the captured cross-correlation.

Extensive theoretical and empirical analyses were conducted to evaluate the performance of Orion and Wadjet using real and synthetic datasets. The evaluation results show that both algorithms have better accuracy and execution time than the state of art techniques when applied to homogeneous data stream applications. The results also show that Wadjet is effective in detecting outliers in heterogeneous data streams which cannot be handled by existing algorithms.

Date: Wednesday, **April 17**, 2013

Time: **4:00 PM**

Place: Devon Hall room **226**

Committee Members: Dr. Gruenwald
Dr. Moses
Dr. Radhakrishnan
Dr. Kim
Dr. Dhall