Lightweight Monitoring of Distributed Streams

Arnon Lazerson, Assaf Schuster. Technion I.I.T. {lazerson, assaf}@cs.technion.ac.il, Daniel Keren. Haifa University. dkeren@cs.haifa.ac.il

Background

Large-scale monitoring applications rely on continuous tracking of complex queries over distributed data streams. Effective distributed stream processing solutions must be:
• Space efficient
• Communication efficient
• Computation efficient

Motivation

Examples for monitoring a function over the average
Distributed Graph

Distributed Monitoring Model

Distributed streams $S_k$ continuously update the local vectors $v_i$. The coordinator $G$ must issue an alert when the global condition $f(\sum v_i) \leq T$ is breached.

Prior Work

• Geometric Monitoring (Sharfman et al) achieved state-of-the-art results in communication reduction.
• However, it places heavy computational burden at the nodes.
• This is a big issue when monitoring rapidly changing data streams.

The Problem

How to define LOCAL conditions at the nodes, such that if they hold, it is guaranteed that the GLOBAL threshold condition holds?

For example, you computed an SVM classifier over distributed data, and then the data changed. Now, you want to locally determine if it’s necessary to recompute the model.

A New Approach

Work with convex functions
• For a convex function $f$, if $f(v_i) \leq T$ holds at every node, it also holds that $f(\sum\frac{v_i}{k}) \leq T$
• Monitoring $f$ (from above) is trivial – just monitor its value at every node

Non-Convex Functions

To monitor a non-convex function – tightly bound the monitored function with a larger convex function (CB, for convex/concave bound), and monitor the CB.

Optimal Bounds

• If $f$ is convex, the optimal bound is of course $f$
• If $f$ is concave, the optimal bound at $p$ is the tangent plane at $t$
• If $f$ is neither, for example $f = x^2 - y^2$, $p = (0, 0)$, the “optimal” bound is $f = x^2$

Real Life Functions

Our method was successfully applied to monitor four important, “real-life” functions:
PCA Score, Cosine Similarity, Inner-Product and the Pearson Correlation Coefficient (PCC)

Example – PCC Bounds

Convex upper bound

Evaluation

Runtime

CB’s runtime is orders of magnitude better
Note: Log scale

Communication reduction
CB was better in all the scenarios we tested.