Scaling Geometric Monitoring Over Distributed Streams

Alexandros D. Keros

June 23, 2016

Supervised by: Prof. V.Samoladas

Table of contents

Introduction

Data Stream Systems[1]

- **Data streams**: Continuous, high volume, size unbound, violative, probably distributed
- *Pull paradigm*
- Centralizing and/or polling $\rightarrow$ prohibitive in terms of communication overhead
- Examples: telecommunication, sensor networks

[1] Brian Babcock et al. "Models and Issues in Data Stream Systems". In: *21st ACM SIGMOD-SIGACT-SIGART*. PODS '02. 2002.

The Geometric Monitoring Method[2]

- Threshold monitoring
- Nodes communicate when needed
  - Local constraints
  - Violation resolution (*false alarms*)
- Arbitrary function monitoring
- Tight accuracy bounds
- A promising framework for *distributed data stream monitoring*

[2] Izchak Sharfman, Assaf Schuster, and Daniel Keren. "A Geometric Approach to Monitoring Threshold Functions over Distributed Data Streams". In: *2006 ACM SIGMOD ICMD*. SIGMOD '06. 2006.

Motivation

Problems:

- increasing node population
- data volume
- data dimensionality
- arbitrary functions
- **communication - accuracy tradeoff**

Need for:

- scalability warranties
- tight accuracy bounds
- incremental/real-time operation
- **Minimize communication while retaining accuracy bounds**

Motivation

Problems:

- increasing node population
- data volume
- data dimensionality
- arbitrary functions
- **communication - accuracy tradeoff**

Need for:

- scalability warranties
- tight accuracy bounds
- incremental/real-time operation
- **Minimize communication while retaining accuracy bounds**

Motivation

Problems:

- increasing node population
- data volume
- data dimensionality
- arbitrary functions
- **communication - accuracy tradeoff**

Need for:

- scalability warranties
- tight accuracy bounds
- incremental/real-time operation
- **Minimize communication while retaining accuracy bounds**

Motivation

Problems:

- increasing node population
- data volume
- data dimensionality
- arbitrary functions
- **communication - accuracy tradeoff**

Need for:

- scalability warranties
- tight accuracy bounds
- incremental/real-time operation
- **Minimize communication while retaining accuracy bounds**

Contributions

Expand the *geometric monitoring method*:

- heuristic method for violation resolution
- distance-based hierarchical node clustering[3]
- throughout method evaluation on synthetic and real-world datasets

[3] Daniel Keren et al. "Geometric Monitoring of Heterogeneous Streams." In: *IEEE Trans. Knowl. Data Eng.* (2014).

Contributions

Expand the *geometric monitoring method*:

- heuristic method for violation resolution
- distance-based hierarchical node clustering[3]
- throughout method evaluation on synthetic and real-world datasets

[3] Daniel Keren et al. "Geometric Monitoring of Heterogeneous Streams." In: *IEEE Trans. Knowl. Data Eng.* (2014).

Contributions

Expand the *geometric monitoring method*:

- heuristic method for violation resolution
- distance-based hierarchical node clustering[3]
- throughout method evaluation on synthetic and real-world datasets

[3] Daniel Keren et al. "Geometric Monitoring of Heterogeneous Streams." In: *IEEE Trans. Knowl. Data Eng.* (2014).

Geometric Threshold Monitoring

System Architecture

Computational Model

Computational Model

Balancing Process

Geometric Interpretation

Convexity Property

Geometric Interpretation

Local Constraints

Protocol

Decentralized Algorithm

Protocol

Centralized Algorithm

Multi-objective Optimization

Non-linear Constraint Optimization

Primal Descent

Feasible Directions

SQP

The Savitzky-Golay Filter

Maximum Weight Matching

The Primal-Dual Method

Related Work

Problem Formulation

The Geometric Monitoring Framework

The Distance-based Hierarchical Clustering

The Idea

The Distance-based Hierarchical Clustering

The Weight Function

The Distance-based Hierarchical Clustering

The Algorithm

The Heuristic Balancing

The Idea

The Heuristic Balancing

The Optimizing Function

The Heuristic Balancing

The Function Formulation

The Heuristic Balancing

The Algorithm

An Nested Optimization Problem

Velocity and Acceleration Estimation via SG Filtering

Implementation Challenges

Synthetic Data

Real-world Data

Notation

RAND, DIST, DISTR Comparison

GM, HM Comparison

GM, HDM Comparison

Synthetic Data Monitoring

GM, HDM Comparison

Air Pollution Monitoring

Summary & Concluding Remarks

Future Work

The end
Questions?