Efficient Threshold Monitoring for Distributed Probabilistic Data
Mingwang Tang, Feifei Li, Jeff M. Phillips, Jeffrey Jestes

Introduction

- Distributed threshold monitoring (DTM) problem concerns about monitoring distributed data and generating an alarm when a user specified threshold constraint is violated.
- Challenges of DTM problem: Uncertainty is inherently introduced in many distributed environments.
  - Data integration produces fuzzy matches
  - Imprecise reading, noisy data in sensor field
  - Multiple reading of the same data
- The Shipboard Automated Meteorological and Oceanographic System (SAMOS)

Preliminary

- Efficient Constraint Monitoring Using Adaptive Thresholds [ICDE08], which deals with DTM problem \((\sum_{i=1}^{n} x_i \leq T)\) in deterministic data.
  - Install local filter \(T_i\) at each client. If \(x_i \leq T_i\), at each client \(c_i\) and \(\sum_{i=1}^{n} T_i \leq T\), \(H\) is sure that \(\sum_{i=1}^{n} x_i \leq T\).
  - If some \(x_i > T_i\) (local alarm), \(H\) check \(\sum_{i=1}^{n} T_i + x_i \leq T\).
  - If \(\sum_{i=1}^{n} T_i + x_i > T\) or \(\sum_{i=1}^{n} x_i > T\) (global poll), \(H\) asks for \(x_i\) from each client and checks \(\sum_{i=1}^{n} x_i \leq T\).

Problem Formulation (PDTM)

Given \(\gamma \in \mathbb{R}^+\) and \(\delta \in [0, 1]\), let \(Y_t = \sum_{i=1}^{n} X_{i,t}\) for \(t = 1, \ldots, T\). The goal is to raise an alarm at \(H\), whenever for any \(t \in [1, T]\) \(Pr[|Y_t| > \gamma] > \delta\). In PDTM problem, our concern is to reduce the communication cost (messages and bytes) and cpu cost.

Exact Methods

- Challenge: A naive way of compute \(Pr[|Y| > \gamma]\) exactly at \(H\) is in \(O(n^2)\) time, where \(n\) is the number of clients and \(n\) is the maximum number of \(<\text{score, probability}>\) pairs in \(X_t\).
- A better way in \(O(n \log n + \log \delta)\) is to compute \(Y_{1,2}, \ldots, Y_{n,2} + \ldots + y\) separately; then, use the cdf (cumulative distribution function) of \(Y_{n,2} + \ldots + y\). We dub this method as ExactD.
- Given that \(E(Y) = \sum_{i=1}^{n} E(X_i)\), each client \(c_i\) only sends \(E(X_i)\), then \(H\) can check if \(E(Y) < \delta\). Then we can leverage on the adaptive threshold algorithm for DTM.

Baseline Method (Adaptive)

- Markov’s inequality gives an upper bound: \(Pr[|Y| > \gamma] \leq \frac{\gamma}{E(Y)}\).
- Given that \(E(Y) = \sum_{i=1}^{n} E(X_i)\), each client \(c_i\) only sends \(E(X_i)\), then \(H\) can check if \(E(Y) < \delta\). Then we can leverage on the adaptive threshold algorithm for DTM.

Improved Algorithm (Adaptive)

I One-sided Chebyshev’s inequality:
\[Pr[|Y| > \gamma] < Pr[|Y| \geq \gamma] \leq \frac{\gamma^2}{E(Y)}\]
\[Pr[|Y| > \gamma] = 1 - Pr[|Y| \leq \gamma] > 1 - \frac{\gamma^2}{E(Y)}\]

II The general Chernoff bound and the moment-generating function
\[Pr[|Y| > \gamma] \leq e^{-\gamma \cdot \phi(\gamma)}\]

Applicable Methods

- Random Distributed <Sample (DDTS): \(H\) gets a sample estimate \(y = \sum_{i=1}^{n} X_i\) from the distribution of \(Y\).
- Deterministic Distributed <Sample (DDTS): \(c_i\) selects \(g_i/s\) evenly spaced points from \(X_i\).
- A randomized improvement of DTS (DDTS): random shift version of DTS with \(\phi(\gamma) = \frac{\sqrt{2 \log \delta}}{\gamma}\) evenly spaced points from \(X_i\).

Experiments

- Respose time
- Precision and recall
- Messages
- Bytes

Default Experimental Parameters

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Default Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tau)</td>
<td>Grouping interval number of clients</td>
<td>100</td>
</tr>
<tr>
<td>(\beta)</td>
<td>Probability threshold</td>
<td>0.7</td>
</tr>
<tr>
<td>(\theta)</td>
<td>Score threshold</td>
<td>0.9</td>
</tr>
<tr>
<td>(n)</td>
<td>Sample size/clients</td>
<td>30</td>
</tr>
</tbody>
</table>

Efficient Algorithm for Distributed Probabilistic Data