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

Introduction

- Distributed Threshold Monitoring (DTM):

  \[ H \uparrow \sum_{i=1}^{n} x_i > 240 \]

- The Shipboard Automated Meteorological and Oceanographic System (SAMOS)

  \[ I \text{adaptive}\] grouping interval

  \[ I \text{adaptive}\] improved

  \[ \sum_{i=1}^{n} \] sample size per client

  \[ 0.6 \] probability threshold

  \[ 0.9 \] applications

  \[ 2000 \] \[ 3000 \] \[ x \]

  \[ 3000 \] \[ 2500 \] \[ 2000 \]


Distributed Probabilistic Threshold Monitoring (DPTM)

- Improved

  \[ H \uparrow \Pr[Y = \sum_{i=1}^{n} X_i > \gamma] > \delta \]

- The lowerbound (upperbound) is a function of some deterministic monitoring instances.

- When derived deterministic monitoring instances fail to make a decision, still expensive to compute Y even with all \( X_i \)’s → use sampling methods!

Baseline Method Based on Markov bound (Madaptive)

- Markov’s inequality:

  \[ \Pr[Y > \gamma] \leq \frac{E(Y)}{\gamma} \]

  \[ H \text{ can check if } \frac{E(Y)}{\gamma} < \delta. \]

  \[ E(Y) = \sum_{i=1}^{n} E(X_i) < \gamma \delta \]

Improved Method

I One-sided Chebyshev’s inequality:

\[ \Pr[Y > \gamma] \leq \frac{1}{\gamma^2} Var(Y) \]

\[ \Pr[Y > \gamma] > 1 - \frac{1}{\gamma^2} \frac{Var(Y)}{E(Y)^2} \]

II The Chernoff bound using the moment-generating function.

\[ M(\beta) = E(e^{\beta Y}) = \prod_{i=1}^{n} M_i(\beta) \text{ for any } \beta \in R \]

\[ \text{for any } \beta_1 > 0 \text{ and } \beta_2 < 0: \]

\[ \sum_{i=1}^{n} \ln M_i(\beta_1) \leq \ln \delta + \beta_1 \gamma \]

\[ \sum_{i=1}^{n} \ln M_i(\beta_2) \leq \ln(1 - \delta) + \beta_2 \gamma \]

- A counter and alarm instances is maintained in each period of \( k \) time instances.

- Periodically decide which monitoring instance to run and set the optimal value of \( \beta_1 \) and \( \beta_2 \)

Random Distributed \( \epsilon \)-Sample (RD\( \epsilon \)S)

- \( H \) asks for a random sample \( x_i \) from each client w.r.t. the distribution of \( X_i \)

- Repeating this sampling \( k = O(\frac{1}{\epsilon} \ln \frac{1}{\epsilon}) \) times.

- \( \Pr[Pr[\hat{Y} > \gamma] - Pr[Y > \gamma]] \leq \epsilon \geq 1 - \phi \) using \( O(\frac{1}{\epsilon^2} \ln \frac{1}{\epsilon}) \) bytes.

Default Experimental Parameters

- Number of clients probability threshold

- Score threshold

- Sample size per client

Datasets

- Real datasets (11.8 million records) from the SAMOS project.

- Each record contain four measurements: WD, WS, SS, TEM, which leads to four single probabilistic attribute datasets.

Experiments

- Response time

- Precision and recall

- Number of messages

- Number of bytes

- Performance of all methods