Anomaly Detection in a Mobile Communication Network

Alec Pawling, Nitesh V. Chawla, and Greg Madey

University of Notre Dame

Overview

We present a technique that uses **hybrid clustering** in conjunction with **statistical process control** to handle **concept drift** in a **data stream.**

Outline

- Motivation
- Background
  - Data streams
  - Concept drift
  - Statistical process control
- Related work
- Hybrid clustering for streams
- Setup
- Results
- Conclusion

Motivation

Application

- *Detection and Alert System* component of WIPER Emergency Response System [Schoenharl *et al.*, 2006], [Madey, *et al.*, 2006]
  - Detect and report anomalies in network usage
  - Notify *Simulation and Prediction System*

Difficulties

- Massive volume of data
- Dynamic system

Data Streams

- Data can only be read once (due to volume)
- Order of data cannot be manipulated
- Often, if the underlying process is stationary, anomaly detection is straightforward
- If the underlying process is dynamic, the problem is difficult

Concept Drift

- Change in process that generates the data stream over time
- May or may not be periodic

Concept Drift

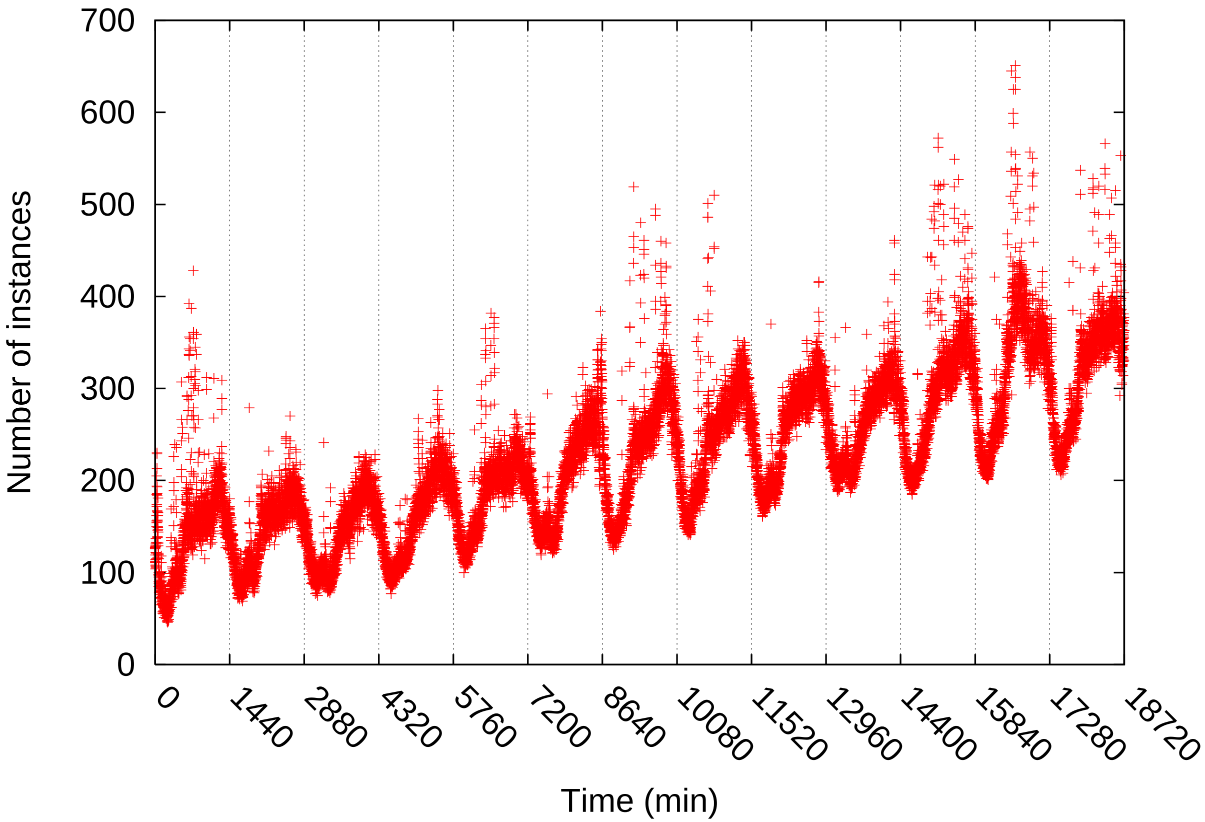

Figure 1: GPRS usage over 12 days

Statistical Process Control

Distinguish between random and assignable variation: threshold is $\mu \pm l\omega$.

- Random variation
  - High probability, little effect on process output
- Assignable variation
  - Low probability, significant effect on process output
  - Change in underlying process

Statistical Process Control

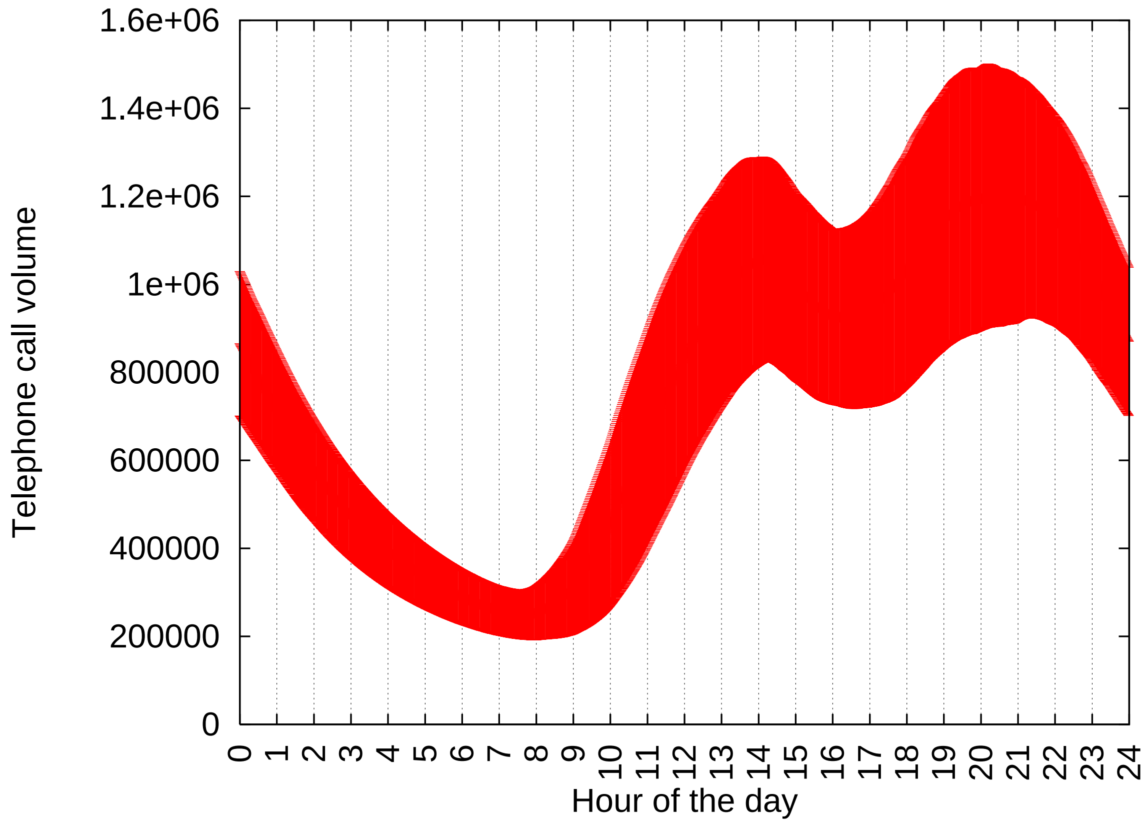

Figure 2: Range of random variance

Related Work

Intrusion detection (Portnoy, 2001)

- Identify intrusions in an unlabeled data set using leader clustering.
- The leader algorithm (Hartigan 1975)
  - Let d be a distance threshold.
  - Let the first instance assigned to cluster C_i be the defining instance, $\mathbf{c}_i$
  - For each instance $\mathbf{x}$
    - Find the closest cluster, C_j
    - If $\text{dist}(\mathbf{x}, \mathbf{c}_i) < d$, add $\mathbf{c}$ to C_i
    - Otherwise, create a new cluster with the defining instance $\mathbf{x}$.

Related Work

Problem:

- Uses z-score normalization to allow for arbitrary data distribution:

$$v_i' = \frac{v_i - \bar{v}_i}{\sigma_i}$$

- This is not possible in one pass

Related Work

Hybrid clustering algorithms, (Cheu *et al.*, 2004)

1. Cluster to reduce the data set
2. Produce final clusters

Hybrid Algorithm for Streams

1. Establish clusters with some minimum number of instances using a partitional or hierarchical algorithm
2. Incrementally update cluster center and standard deviations using a variation on the leader algorithm.

Setup

Data set

- Feature vector consists of timestamp and number of instances of 5 services
- One example for each minute of a 12 day period (18721 examples)

Clustering Algorithms

- Expectation Maximization — Weka, cross-validation to determine number of clusters
- Leader
- Hybrid for streams: (1) k-means, (2) modified leader

Results

Hybrid algorithm

- Small clusters compared to EM
- Little consistency in detected outliers among different thresholds or values of k

Leader algorithm

- More consistency in anomaly detection

Conclusion

- Algorithms using random values may be a bad idea
- Algorithms requiring only threshold parameter seem promising

Future work

- Hierarchical clustering to establish clusters
- Examine further how the number of clusters grows over time

Acknowledgments

This work is supported, in part, by the National Science Foundation, the DDDAS Program, under Grant No. CNS_050348.