Detecting Outliers in Data with Correlated Measures

Yu-Hsuan Kuo

Computer Science & Engineering
Penn State University

October 23, 2018

Joint work with Zhenhui Li and Daniel Kifer

Outline

Motivation

- In large-scale sensor datasets, there could be a significant amount of outliers due to sensor malfunction or human operation faults.

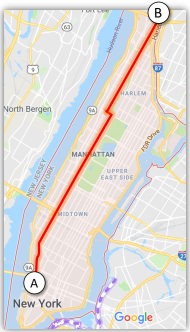


Figure: long moving distance but unreasonably low trip fare

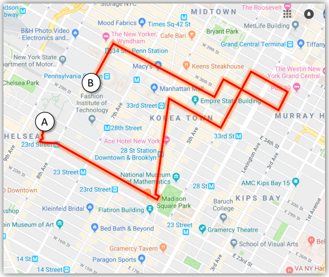
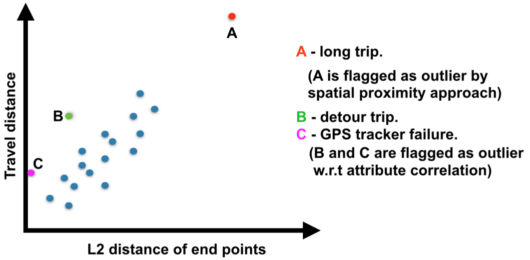


Figure: short L2 distance between pickup (A) and dropoff (B) but long trip distance

- Such outliers in the original datasets can break effective travel time estimation methods [WKKL16].

Contextual Outlier [SWJR07]

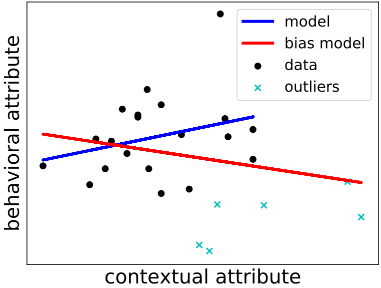
- Typical outlier detection defines a sample as an outlier if it significantly deviates from other data samples. $\Rightarrow$ not apply in our case.



- Contextual outlier detection: use the correlation between contextual attributes and behavioral attributes [SWJR07, HH15, LP16].
- We detect outliers based on empirical correlations of attributes. (e.g., trip time and trip distance)
- Anomaly: attributes of a data sample significantly deviate from expected correlations.

Related Work

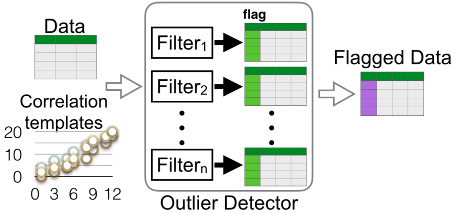
- One problem with contextual outlier detection [SWJR07, HH15, LP16] is that outliers can bias a model learned from noisy data.



- Clean data is almost not available $\Rightarrow$ contextual outlier detector trained on noisy data.
- Our solution: a robust regression model that explicitly considers outliers.

System Overview

- Input: Data & Correlation templates (j, S) where j is behavior attribute and S is a set of contextual attributes
- Output: flagged suspicious records.
- A **filter**:
  1. take correlation template (j, S) and learn, for each record $\vec{z}_i$, how to predict behavior attribute $\vec{z}_i[j]$ from contextual attributes $\vec{z}_i[s]$ for $s \in S$.
  2. assign an outlier score t_i to every record.
  3. provide an estimate for the total number of outliers.
- A record is marked as outlier if at least one filter marks it as an outlier.



Outline

Mixture Model

- For a correlation (j, S), let y_i be behavioral attribute value and $\vec{x}_i$ be the vector of contextual attribute values in S.
- Learn a model that can predict y_i from the attributes $\vec{x}_i$.

$$y_i = \vec{w} \cdot \vec{x}_i + \epsilon_i$$

- Model the prediction error: a mixture of light-tailed distributions (for non-outliers) and heavy-tailed distributions (for outliers).
- Assume there is a probability p that a data point is an outlier $\Rightarrow$ Noise distribution ϵ_i for record i: with prob. $1 - p$ it is a Gaussian, and with prob. p it is a Cauchy random variable.

Likelihood Function

1. A zero mean Gaussian with unknown variance σ^2 has probability density

$$f_G(\epsilon_i; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\epsilon_i^2}{2\sigma^2})$$

2. Cauchy distribution with scale parameter b is a heavy-tailed distribution with undefined mean and variance $\Rightarrow$ ideal for modeling outliers.
   - A sample ϵ_i from this distribution: first sampling a value τ_i from the Gamma(0.5, b) distribution then sampling ϵ_i from the Gaussian(0, $1/\tau_i$) distribution [BL09]:

$$f_C(\epsilon_i, \tau_i; b) = \frac{b^{0.5}}{\Gamma(0.5)} {\tau_i}^{0.5-1} e^{-b\tau_i} \frac{\sqrt{\tau_i}}{\sqrt{2\pi}} \exp(-\frac{\tau_i \epsilon_i^2}{2})$$

Likelihood Function (Cont.)

3. Latent indicator χ_i: where the error of contextual attribute $\vec{x}_i$ comes (i.e. from Cauchy or Gaussian)
4. With the model parameters $\vec{w}$, unknown noise parameters σ^2 (variance of non-outliers), p (outlier probability), b (scale parameter of outlier distribution), the likelihood function is

$$
\begin{aligned}
&L(\vec{w}, \sigma^2, p, b, \vec{\chi}, \vec{\tau}) \\
&= \prod_{i=1}^{n} \left[(1-p) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i - \vec{w} \cdot \vec{x}_i)^2}{2\sigma^2}) \right]^{1-\chi_i} \times \\
&\left[p \frac{b^{0.5}}{\Gamma(0.5)} {\tau_i}^{0.5-1} e^{-b\tau_i} \frac{\sqrt{\tau_i}}{\sqrt{2\pi}} \exp(-\frac{\tau_i (y_i - \vec{w} \cdot \vec{x}_i)^2}{2}) \right]^{\chi_i}
\end{aligned}
$$

Parameters Learning

- EM algorithm [DLR77] to solve the likelihood function L.
- E step:
  - parameter τ_i of Cauchy density
  - estimated probability that it is an outlier t_i (i.e. expected value of χ_i)
  - scale parameter b
- M step:
  - estimated fraction of outliers p
  - the variance of non-outliers σ^2
  - model coefficients $\vec{w}$
- Outlier labeling: every filter model assigns to every record i a score t_i. It then labels a record an outlier if it has one of the top K values of t_i where $K = \lfloor \sum_{i=1}^{n} t_i \rfloor \approx p \times$ total number of records n.

Outline

1. Introduction: Detecting Outliers in Big Data
2. Outlier Data Modeling
3. Experimental Results

Datasets

Use 4 datasets

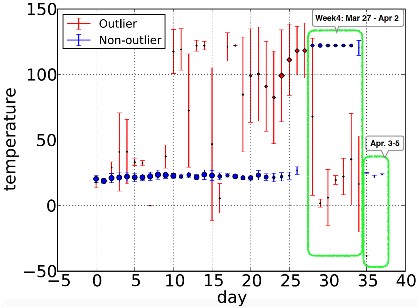
1. **NYC Taxi**: A large-scale public New York City taxi dataset is collected from more than 14,000 taxis, which contains $173,179,771$ taxi trips in 2013.
2. **Intel Lab Sensor**: A public Intel sensor dataset containing a log of about 2.3 million readings from 54 sensors deployed in the lab.
3. **ElNino**: A dataset contains 93,935 records. These readings are collected from buoys positioned around equatorial Pacific.
4. **Houses**: A dataset with 20,640 observations on the housing in California.

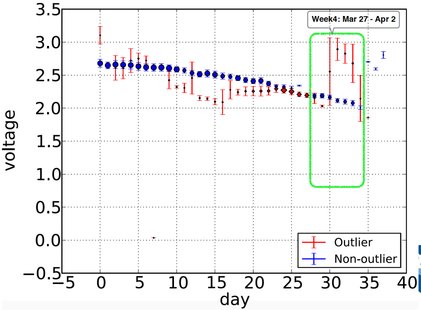
Baselines

- **Density-based method**. A widely referenced density-based algorithm LOF [BKNS00] outlier mining.
- **Distance-based method**. A recent distance-based outlier detection algorithm with sampling [SB13].
- **OLS**. The linear regression with ordinary least square estimation.
- **GBT**. The gradient boosting tree regression model [Fri01].
- **CAD**. Conditional Anomaly Detection [SWJR07].
- **ROCOD**. Robust Contextual Outlier Detection [LP16].

Intel Sensor Data Results

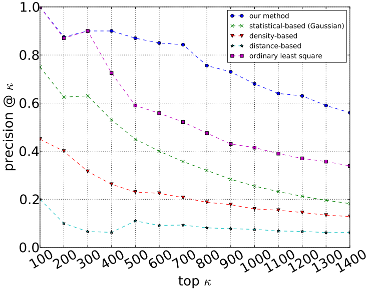
- No ground truth ⇒ validate with findings in the Scorpion system [WM13], & case study.
- 57.2% of flagged outliers are anomalous temperature reading
- Observed a general sensor's malfunction pattern as it is unlikely to be real temperature in the lab.
- A decreasing trend in voltage for this batch of sensors.





NYC Taxi Data Results

- We designed a human labeling system for experienced taxi riders to determine outlier trips.
- Evaluation metric: Precision $@\kappa = \dfrac{\#\text{ trips whose rank} \leq \kappa \text{ and label} = \text{Outlier}}{\kappa}$



- Our top outlier trips are mainly from device error (e.g. unreasonable trip distance, trip fare < min fare, trips with GPS failure)

Experiments on Synthetic Outlier Data

- We inject synthetic outliers into Elnino and Houses datasets.
- Perturbation scheme: inject q % of outliers into N data samples.
  - randomly select $q \times N$ records $\vec{z_i} = (\vec{x_i}, y_i)$ to be perturbed.
  - a random number from $(0, \alpha)$ is added up to target attribute y_i as ${y_i}'$.
  - add new sample $\vec{z}\,' = (\vec{x_i}, {y_i}')$ as outlier.
- Evaluation metric: the Area Under the Curve (AUC) of the Precision-Recall curve.

Synthetic Outlier Results - Perturb Behavioral Attributes

- Our outlier detector consistently performs the best when more outliers are involved.

Table: PR AUC w.r.t different fractions of synthetic outliers in behavioral attribute

	Elnino				
method	q=0.01	q=0.03	q=0.05	q=0.1	q=0.15
Doc	**0.96**	**0.97**	**0.98**	**0.98**	**0.98**
ROCOD (non-linear)	0.73	0.73	0.74	0.73	0.72
CAD	0.80	0.84	0.86	0.85	0.88
OLS	**0.96**	0.95	0.95	0.92	0.90
GBT	**0.96**	0.95	0.95	0.92	0.90
distance-based	0.81	0.74	0.77	0.83	0.60
density-based	0.21	0.38	0.45	0.38	0.34

Synthetic Outlier Results - Perturb Behavioral Attributes (Cont.)

Table: PR AUC w.r.t different fractions of synthetic outliers in behavioral attribute

	Houses				
method	q=0.01	q=0.03	q=0.05	q=0.1	q=0.15
Doc	**0.93**	**0.92**	**0.93**	**0.95**	**0.96**
ROCOD (non-linear)	0.50	0.49	0.50	0.49	0.50
CAD	0.58	0.67	0.68	0.72	0.75
OLS	0.92	0.91	0.92	0.91	0.91
GBT	**0.93**	0.91	0.92	0.91	0.91
distance-based	0.76	0.19	0.57	0.4	0.39
density-based	0.84	0.58	0.46	0.53	0.58

Synthetic Outlier Results - Perturb Contextual Attributes

- A small fraction of outliers in contextual attribute hurts the performance considerably for the other methods.
- Our method is robust and resistant to the fraction of outliters.

Table: PR AUC w.r.t different fractions of synthetic outliers in contextual attribute

	Elnino				
method	q=0.005	q=0.01	q=0.03	q=0.05	q=0.07
Doc	**0.97**	**0.95**	**0.97**	**0.98**	**0.98**
ROCOD (non-linear)	0.01	0.01	0.02	0.02	0.03
CAD	0.80	0.83	0.86	0.88	0.87
OLS	0.92	0.86	0.68	0.45	0.32
GBT	0.11	0.15	0.28	0.37	0.40
distance-based	0.88	0.74	0.81	0.50	0.83
density-based	0.08	0.07	0.08	0.09	0.10

Synthetic Outlier Results - Perturb Contextual Attributes (Cont.)

Table: PR AUC w.r.t different fractions of synthetic outliers in contextual attribute

	Houses				
method	q=0.005	q=0.01	q=0.03	q=0.05	q=0.07
Doc	**0.86**	**0.80**	**0.88**	**0.88**	**0.91**
ROCOD (non-linear)	0.03	0.01	0.02	0.04	0.05
CAD	0.51	0.54	0.56	0.61	0.63
OLS	0.84	0.75	0.71	0.59	0.50
GBT	0.04	0.04	0.08	0.11	0.15
distance-based	0.54	0.73	0.22	0.20	0.42
density-based	0.01	0.01	0.03	0.04	0.06

Synthetic Outlier Results - Degree of Outlierness

- As α increases, larger magnitude of noise will have more chance to be added to the original value.
- Our performance increased as more extreme outliers are added.

Table: PR AUC w.r.t degree of outlierness α in contextual attribute

	Elnino				
method	$\alpha = 20$	$\alpha = 30$	$\alpha = 50$	$\alpha = 100$	$\alpha = 300$
Doc	**0.91**	**0.94**	**0.95**	**0.98**	**0.99**
ROCOD (non-linear)	0.01	0.01	0.01	0.01	0.01
CAD	0.78	0.8	0.83	0.87	0.93
OLS	0.88	0.89	0.86	0.85	0.73
GBT	0.17	0.17	0.15	0.17	0.17
distance-based	0.21	0.79	0.74	0.88	0.91
density-based	0.13	0.10	0.07	0.05	0.04

Synthetic Outlier Results - Degree of Outlierness (Cont.)

Table: PR AUC w.r.t degree of outlierness α in contextual attribute

	Houses				
method	$\alpha = 30$	$\alpha = 50$	$\alpha = 100$	$\alpha = 300$	$\alpha = 500$
Doc	**0.75**	**0.8**	**0.94**	**0.97**	**0.99**
ROCOD (non-linear)	0.01	0.01	0.01	0.01	0.01
CAD	0.37	0.54	0.58	0.74	0.85
OLS	0.72	0.75	0.87	0.86	0.83
GBT	0.04	0.04	0.03	0.02	0.01
distance-based	0.14	0.73	0.79	0.85	0.80
density-based	0.01	0.01	0.01	0.02	0.05

Summary

- We develop a system to detect outliers by correlations between measurements.
- It is a robust model as compared to the existing algorithms built on all the data records where their model parameters are skewed by outliers.
- We compare our approach against traditional outlier detectors, contextual outlier detectors and regression models. Our method significantly outperformed competing methods and continues to perform well even in extremely noisy datasets.

Questions?

References I

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander.
Lof: identifying density-based local outliers.
In SIGMOD, 2000.

Narayanaswamy Balakrishnan and Chin-Diew Lai.
Continuous bivariate distributions.
Springer Science & Business Media, 2009.

A. P. Dempster, N. M. Laird, and D. B. Rubin.
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society: Series B, 39:1–38, 1977.

Jerome H Friedman.
Greedy function approximation: a gradient boosting machine.
Annals of statistics, pages 1189–1232, 2001.

References II

Charmgil Hong and Milos Hauskrecht.
Mcode: Multivariate conditional outlier detection.
arXiv preprint arXiv:1505.04097, 2015.

Jiongqian Liang and Srinivasan Parthasarathy.
Robust contextual outlier detection: Where context meets sparsity.
In CIKM. ACM, 2016.

Mahito Sugiyama and Karsten Borgwardt.
Rapid distance-based outlier detection via sampling.
In NIPS, 2013.

Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka.
Conditional anomaly detection.
ICDE, 2007.

References III

- Hongjian Wang, Yu-Hsuan Kuo, Daniel Kifer, and Zhenhui Li.
  A simple baseline for travel time estimation using large-scale trip data.
  In SIGSPATIAL. ACM, 2016.

- Eugene Wu and Samuel Madden.
  Scorpion: Explaining away outliers in aggregate queries.
  In VLDB Journal, 2013.