Semantic-aware anomaly detection in real world parking data

Arnamoy Bhattacharyya, Weihan Wang, Cristine Tang, Cristiana Amza

Abstract

In this work, we introduce and experimentally evaluate a novel approach for anomaly detection in smart car parking applications. We attach semantics to the raw parking data collected from sensors of parking lots to detect anomalous patterns in them. Attaching semantics on top of raw data helps reduce the processing time and also provides the error checker a distinct context to look into potential problems.

1. Introduction

TODO

2. A Semantic Aware Anomaly Detection Approach

The data we are analyzing has different car occupancy metrics for parking data collected from the parking lots at various locations. There are two types of car occupancy: contract based cars and transient parkers.

The data we have is for n garages, collected over m hours.

$$contacts = \{con_1, con_2, \ldots, con_n\} \tag{1}$$

$$transients = \{tran_1, tran_2, \ldots, tran_n\} \tag{2}$$

Parking data for each garage i ($i \in n$) has the data for m months and has the following form:

$$con_i = \{mon_1, mon_2, \ldots, mon_m\} \tag{3}$$

$$tran_i = \{mon_1, mon_2, \ldots, mon_m\} \tag{4}$$

Each set of monthly parking data mon_j^i ($i \in n, j \in m$) has d sets of data points, depending on the number of days that month has. The form for monthly data is:

$$mon_j^i = \{day_1, day_2, \ldots, day_d\} \tag{5}$$

ISBN [to be supplied]. . . $15.00

Finally, each set day_k ($k \in d$) has 24 data points, one for each hour of the day.

2.1 Semantics of the parking data

To provide a faster analysis of the huge amount of parking data and to provide meaningful feedback to the data checker after the automatic anomaly detection, we attach semantics on top of the raw time-series parking data. These *modified signals*, with the attached semantics on top are then fed to our anomaly detection methods to find abnormal patterns in them. We define the following *modified signals* that we derive from the raw parking data:

1. **Monthly Peak Occupancy**: The value of a monthly peak occupancy signal mon_{pk} for a garage $garage_i$ for m months is given by the following equation:

$$\{mon_{pk}(m)\} = (\max\{mon_1^i\}, \max\{mon_2^i\}, \ldots, \max\{mon_m^i\}) \tag{6}$$

   We further divide the mon_{pk} signal into two finer semantics: contracts ($mon_{pk_{con}}$) and transients($mon_{pk_{tran}}$). The number of data points in each signal $mon_{pk_{con}}$ and $mon_{pk_{tran}}$ is the number of months under analysis. This semantics help us to differentiate two distinct parking behaviour patterns for the two parking types.

2. **Daily Peak Occupancy**: The value of a daily peak occupancy signal day_{pk} for a garage $garage_i$ for d days is given by the following equation:

$$\{day_{pk}(d)\} = (\max\{day_1^i\}, \max\{day_2^i\}, \ldots, \max\{day_d^i\}) \tag{7}$$

   Where $\{day_d^i\}$ is the set of hourly occupancies for the d-th day for garage i. Similar to the monthyl peak occupancy, we have two finer semantics $day_{pk_{con}}$ and $day_{pk_{tran}}$ for contracts and transient type of parking.

3. **Daily Occupancy**: The daily occupancy is the signal that is comprised of all the 24 datapoints of that given day. Then we compose a signal $day\,(d)$ that is a concatenation

of all the datapoints of d days. This signal has $24\dot{d}$ datapoints. We have day_{con} and mon_{tran} for contract based and transient parkings respectively.

2.2 Types of anomalies

Once we have identified the above semantics (at different granularity levels), we define two different types of anomalies for each semantics. They are the following:

1. **Zero Anomalies**: When the data value is '0' at certain places.
2. **Unusual Anomalies**: When the data values shows unusual values as compared to the other values we are analyzing.

2.3 Anomaly detection methods

We use different methods for detecting anomalies at different semantic granularity level. First we describe those methods. Then we describe how we optimize our anomaly detection algorithm accordingly as more anomalies are being detected on the training data.

1. **Zero Anomaly detection**: We detect if there is any zero present in the signals $mon_{pk_{con}}$, $mon_{pk_{tran}}$, $day_{pk_{con}}$ and $day_{pk_{tran}}$. Once we have detected the zeros on the mentioned signals, we also detect if there are sequences of zeros in those anomalies. A sequence of zeros in monthly peak occupancies specify a severe error. A sequence of zeros in the daily peak anomalies specify an error with medium severity. We do not raise an alarm for zero anomalies in daily occupancy.
2. **Unusual Anomaly detection**: Detecting unusual patterns in data is a non-trivial task. But attaching semantics on top of the raw data helps greatly in detection and defining anomaly categories.
   - Detecting *Global* anomalies: We use an well-known statistical technique used in outlier detection for detecting these kind of anomalies. The method is based on Interquartile Range (IQR). In descriptive statistics, the interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles. Data points that falls beyond $\pm 1.5\dot{I}QR$ are generally defined as outliers (1). But from our experiments, a value of 1.5 was too restrictive. So we chose a value of 3.. Anything beyond $\pm 3\dot{I}QR$ is flagged as anomalies by our method.
   - Detecting *Local* anomalies: The above method can detect outliers for data where there is not much change over time. But if there is a trend of increase and decrease in occupancies over time, there can be a few anomalies which are not detected by the above method. One case is shown in Figure 1. Here there was an increasing trend in the daily peak anomalies. Therefore the anomaly during October (red corcle in figure 1 is not detected by our *global* method as the there are multiple datapoints towards the beginning of the year that will change the nature of the distribution and not flag the anomaly in October. In these scenarios, we have to look at local trends (months around October) to detect anomalies. We use a sliding window based approach to detect these *local* anomalies.

     We define a window size w, calculated by Algorithm 1

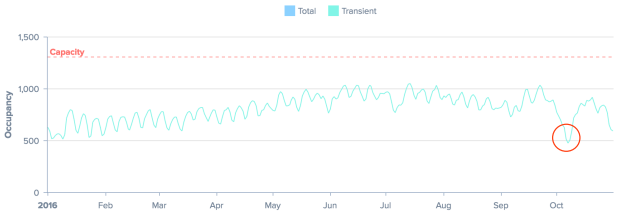

Figure 1: Example of local anomaly.

Input: data_size, max_slides, min_window
Output: window_size
num_slides = 0 ;
window_size = min_window ;
if *data_size <max_slides* **then**
  return window_size ;
while *num_slides <max_slides* **do**
  window_size = data_size - num_slides +1 ;
  num_slides ++ ;
end
return window_size ;

Algorithm 1: Algorithm to determine the sliding window size for detecting *local* anomalies.

     We have a minimum window size min_window. The window size window_size is calculated by a cap on the maximum number of slides, max_slides.

   - Detecting *minor* anomalies: Apart from the other two methods, there can still be anomalies that are not typical outliers but they are derivation from usual *patterns* for the metric of consideration. TODO: Give an example For detecting these anomalies, we use an algorithm called S-H-ESD (ref), or Seasonal Hybrid Extreme Studentized Deviates. This is an extension of a generalized ESD method by adding steps which break down data series into piecewise approximations (a hybrid method) and can account for seasonality in each submodel. Generalized Extreme Studentized Deviates (ESD) is a well established statistical procedure for the detection of outliers where the assumption is

that the inliers are normal distributed. The generalized ESD method introduces a procedure for identifying from 1 to k outliers in a dataset simultaneously thus introducing a robustness for the particular number of outliers present.

The ESD procedure computes statistics $R_1, \ldots, R_k$ for k data points as the extreme studentized deviates. The first statistic is the largest deviation from the mean, $R_1 = \max_i \frac{|x_i - \bar{x}|}{s}, \quad s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$, and the next statistic is calculated on the reduced sample size after removing the sample with the largest deviation, and so on for the subsequent statistics. Critical values λ_i for each test statistic are determined based on a transformed t distribution for a specified confidence level (2). The decision rule is given as follows: if all of the test statistics are lower than the critical values, then there are no outliers. If any of the test statistics are greater than the critical value, then the largest number of points such that the associated test statistic is greater than the critical value are removed as outliers.

References

[1] W. C. Navidi. *Statistics for engineers and scientists*, volume 1. McGraw-Hill New York, 2006.

[2] B. Rosner. Percentage points for a generalized esd many-outlier procedure. *Technometrics*, 25(2):165–172, 1983.