COMPSCI 762 Tutorial 10

Tutorial 11 – Anomaly Detection and Data Stream Mining

Luke Chang

June 2021

The University of Auckland

Topics

Anomaly Detection

Anomaly Detection

Types of Anomaly

- **Global outlier (Point anomaly):** deviates significantly from the rest of the data set. The simplest type of outliers.
- **Contextual outlier (Conditional outlier):** deviates significantly with respect to a specific context of the object.
  - **Contextual attributes:** define the object's context.
  - **Behavioral attributes:** define the object's characteristics, and are used to evaluate whether the object is an outlier in the context.

Example

A temperature sensor measures 4°C in May. It is a perfectly normal reading in Wellington, but it might be an outlier if the location is New York. The location and the data are **contextual attributes**, and the temperature is a **behavioral attribute**.

- **Collective outliers:** the objects as a whole deviate significantly from the entire data set.

Types of Anomaly

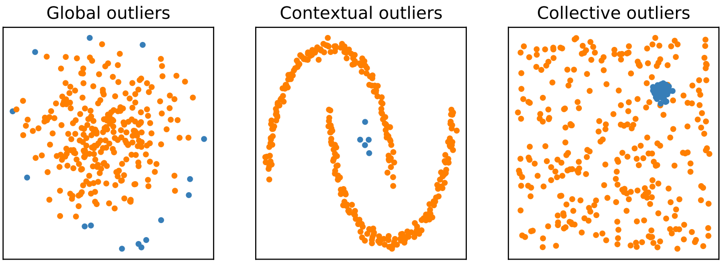
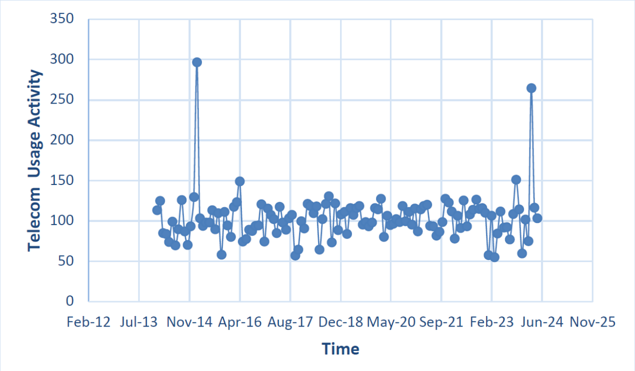


Figure 1: Types of outliers

- **Fig 1a:** Global outliers
- **Fig 1b:** Contextual outliers – Given the dataset has two clusters, each one has a moon shape.
- **Fig 1c:** Blue points are collective outliers because the density of those points is much higher than the rest.

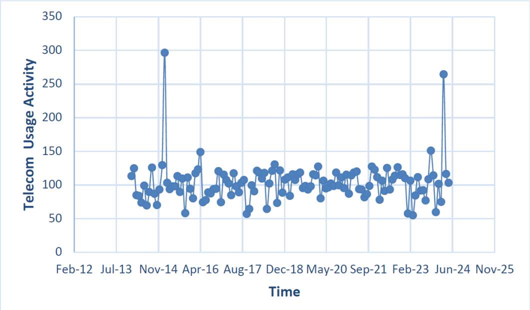
2020 361 Exam Question 4 – Outlier / Anomaly Detection

The following figure shows the monthly telecom usage activity in a specific area across several years. You were tasked to identify whether outliers exist in the data.



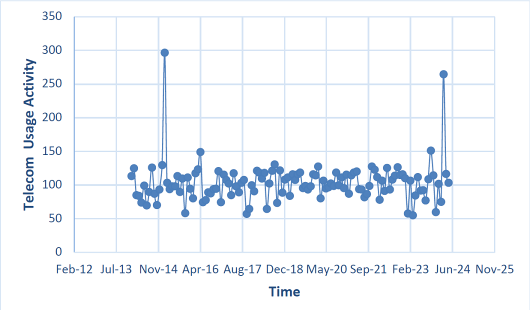
What outlier detection technique would you use to identify whether outliers exist for this case?

2020 361 Exam Question 4 – Outlier / Anomaly Detection



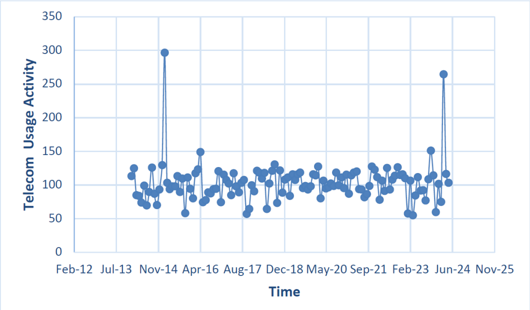
- No obvious periodic pattern;
- No dramatic density changes;
- Unlikely to have contextual outlier and collective outlier;
- We apply the **statistical approach – parametric method** to find global outliers.
- There are multiple ways to solve this problem.

2020 361 Exam Question 4 – Outlier / Anomaly Detection



- **Parametric method:** Assume the samples are drawn from a normal distribution, estimate the maximum likelihood of mean μ and standard deviation σ.
- We know that the $\mu \pm 3\sigma$ region contains 99.7%. According to z-score, if $z = \frac{|x_i - \bar{x}|}{s} > 3$, the sample x_i is generated by the normal distribution is less than $\frac{0.3}{2} = 0.15\%$.
- Note that we do not know the population mean and standard deviation, we only have sample mean and sample standard deviation.

2020 361 Exam Question 4 – Outlier / Anomaly Detection



- Based on observation, each grid contains around 14 points, we have 110 data points in total.
- $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \approx 100$
- $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \approx 20$
- $\frac{|300-100|}{20} = 10 > 3$, an outlier
- $\frac{|270-100|}{20} = 8.5 > 3$, an outlier
- $\frac{|150-100|}{20} = 2.5 < 3$, not an outlier
- $\frac{|55-100|}{20} = 2.25 < 3$, not an outlier

Tutorial Question 1 – Anomaly Detection

Question 1

You are given the following list of 2D data points. [[1; 1]; [1; 2]; [2; 2]; [2; 1]; [3; 3]; [2; 5]; [2; 3]] If you had to select one point to be anomalous, which would you pick? Explain your answer. Please link your explanation to an anomaly detection technique.

There are multiple ways to solve this problem.
Let's use **distance-based outlier detection**.

Tutorial Question 1 – Anomaly Detection

Distance-based outlier detection

	1,1	1,2	2,2	2,1	3,3	2,5	2,3
1,1	0	1	2	1	4	5	3
1,2	1	0	1	2	3	4	2
2,2	2	1	0	1	2	3	1
2,1	1	2	1	0	3	4	2
3,3	4	3	2	3	0	3	1
2,5	5	4	3	4	3	0	2
2,3	3	2	1	2	1	2	0

Table 1: Manhattan Distance Matrix

$$\frac{\|\{o'|\text{dist}(o,o') \leq r\}\|}{\|D\|} \leq \pi$$

We need to define the hyperparameters: r and π.

The number of data points, $\|D\|$, is 7.

Let $r = 2$,

	# of objects within r	Divide by $\|D\|$
1,1	3	0.43
1,2	4	0.57
2,2	5	0.71
2,1	4	0.57
3,3	2	0.29
2,5	1	0.14
2,3	5	0.71

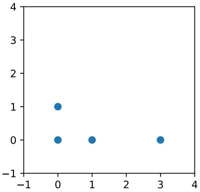
Table 2: The # of neighbours within r

If we select only one point as an outlier, we can set π to any value between 0.14 and 0.29, e.g. 0.15. Therefore, we identify $[2; 5]$ is an outlier.

Tutorial Question 2 – Density-based Approach: Local Outlier Factor (LOF)

Question 2

Consider a set of points (0,0), (1,0), (0,1), (3,0). Calculate the LOF score for the points using Manhattan distance and k is 2.



	a (0,0)	b (1,0)	c (0,1)	d (3,0)
a (0,0)	0	1	1	3
b (1,0)	1	0	2	2
c (0,1)	1	2	0	4
d (3,0)	3	2	4	0

Table 3: Manhattan distance matrix

	$\text{dist}_2(o)$	$N_2(o)$
a (0,0)	1	2
b (1,0)	2	3
c (0,1)	2	2
d (3,0)	3	2

Table 4: Distance between data point o and its k-th nearest neighbour

$$N_k(o) = \{o' | o' \in D, \text{dist}(o, o') \leq \text{dist}_k(o)\}$$

Note that $N_k(o)$ may contain more than k objects, because objects may have same distance.

Tutorial Question 2 – LOF

Let o' be a neighbour of o, to avoid $\text{dist}(o, o')$ is too small, we compute the reachability distance from o' to o:

$$\text{reachdist}_k(o \leftarrow o') = \mathsf{max}\{\text{dist}_k(o), \text{dist}(o, o')\}$$

Note that reachability distance is not symmetric, thus

$$\text{reachdist}_k(o \leftarrow o') \neq \text{reachdist}_k(o' \leftarrow o)$$

Tutorial Question 2 – LOF

$$
\begin{aligned}
\mathrm{reachdist}_2(a \leftarrow b) &= \max\{\mathrm{dist}_2(a), \mathrm{dist}(a,b)\} &&= \max\{1,1\} &&= 1\\
\mathrm{reachdist}_2(a \leftarrow c) &= \max\{\mathrm{dist}_2(a), \mathrm{dist}(a,c)\} &&= \max\{1,1\} &&= 1\\
\mathrm{reachdist}_2(a \leftarrow d) &= \max\{\mathrm{dist}_2(a), \mathrm{dist}(a,d)\} &&= \max\{1,3\} &&= 3\\
\mathrm{reachdist}_2(b \leftarrow a) &= \max\{\mathrm{dist}_2(b), \mathrm{dist}(b,a)\} &&= \max\{2,1\} &&= 2\\
\mathrm{reachdist}_2(b \leftarrow c) &= \max\{\mathrm{dist}_2(b), \mathrm{dist}(b,c)\} &&= \max\{2,2\} &&= 2\\
\mathrm{reachdist}_2(b \leftarrow d) &= \max\{\mathrm{dist}_2(b), \mathrm{dist}(b,d)\} &&= \max\{2,2\} &&= 2\\
\mathrm{reachdist}_2(c \leftarrow a) &= \max\{\mathrm{dist}_2(c), \mathrm{dist}(c,a)\} &&= \max\{2,1\} &&= 2\\
\mathrm{reachdist}_2(c \leftarrow b) &= \max\{\mathrm{dist}_2(c), \mathrm{dist}(c,b)\} &&= \max\{2,2\} &&= 2\\
\mathrm{reachdist}_2(c \leftarrow d) &= \max\{\mathrm{dist}_2(c), \mathrm{dist}(c,d)\} &&= \max\{2,4\} &&= 4\\
\mathrm{reachdist}_2(d \leftarrow a) &= \max\{\mathrm{dist}_2(d), \mathrm{dist}(d,a)\} &&= \max\{3,3\} &&= 3\\
\mathrm{reachdist}_2(d \leftarrow b) &= \max\{\mathrm{dist}_2(d), \mathrm{dist}(d,b)\} &&= \max\{3,2\} &&= 3\\
\mathrm{reachdist}_2(d \leftarrow c) &= \max\{\mathrm{dist}_2(d), \mathrm{dist}(d,c)\} &&= \max\{3,4\} &&= 4
\end{aligned}
$$

$$\mathrm{reachdist}_k(o \leftarrow o') \neq \mathrm{reachdist}_k(o' \leftarrow o)$$

Tutorial Question 2 – LOF

The *Local Reachability Density* (LRD) of an object o is defined as

$$\mathrm{lrd}_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} \mathrm{reachdist}_k(o' \leftarrow o)}$$

$$\begin{aligned}
\mathrm{lrd}_2(a) &= \|N_2(a)\| / (\mathrm{reachdist}_2(b \leftarrow a) + \mathrm{reachdist}_2(c \leftarrow a)) && = 2/(2+2) && = 0.5 \\
\mathrm{lrd}_2(b) &= \|N_2(b)\| / (\mathrm{reachdist}_2(a \leftarrow b) + \mathrm{reachdist}_2(c \leftarrow b) + \mathrm{reachdist}_2(d \leftarrow b)) && = 3/(1+2+3) && = 0.5 \\
\mathrm{lrd}_2(c) &= \|N_2(c)\| / (\mathrm{reachdist}_2(a \leftarrow c) + \mathrm{reachdist}_2(b \leftarrow c)) && = 2/(1+2) && \approx 0.667 \\
\mathrm{lrd}_2(d) &= \|N_2(d)\| / (\mathrm{reachdist}_2(a \leftarrow d) + \mathrm{reachdist}_2(b \leftarrow d)) && = 2/(3+2) && = 0.4
\end{aligned}$$

Tutorial Question 2 – LOF

The *Local Outlier Factor* (LOF) of an object o is

$$\mathrm{LOF}_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{\mathrm{lrd}_k(o')}{\mathrm{lrd}_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} \mathrm{lrd}_k(o') \cdot \sum_{o' \in N_k(o)} \mathrm{reachdist}_k(o' \leftarrow o)$$

$$\begin{aligned}
\mathrm{LOF}_2(a) &= (\mathrm{lrd}_2(b) + \mathrm{lrd}_2(c)) \cdot \textstyle\sum_{o' \in N_k(o)} \mathrm{reachdist}_2(o' \leftarrow a) && = (0.5 + 0.667) \cdot (2 + 2) && = 4.668 \\
\mathrm{LOF}_2(b) &= (\mathrm{lrd}_2(a) + \mathrm{lrd}_2(c) + \mathrm{lrd}_2(d)) \cdot \textstyle\sum_{o' \in N_k(o)} \mathrm{reachdist}_2(o' \leftarrow b) && = (0.5 + 0.667 + 0.4) \cdot (1 + 2 + 3) && = 9.402 \\
\mathrm{LOF}_2(c) &= (\mathrm{lrd}_2(a) + \mathrm{lrd}_2(b)) \cdot \textstyle\sum_{o' \in N_k(o)} \mathrm{reachdist}_2(o' \leftarrow c) && = (0.5 + 0.5) \cdot (1 + 2) && = 3 \\
\mathrm{LOF}_2(d) &= (\mathrm{lrd}_2(a) + \mathrm{lrd}_2(b)) \cdot \textstyle\sum_{o' \in N_k(o)} \mathrm{reachdist}_2(o' \leftarrow d) && = (0.5 + 0.5) \cdot (3 + 2) && = 5
\end{aligned}$$

Tutorial Question 2 – LOF

```python
import numpy as np
from sklearn.neighbors import LocalOutlierFactor

data = np.array([[0,0],[1,0],[0,1],[3,0]])
# The neighbours include itself.
clf = LocalOutlierFactor(n_neighbors=3, metric='manhattan', novelty=False)
pred = clf.fit_predict(data)
print(pred)
# [1 1 1 1]
print(clf.negative_outlier_factor_)
# [-1.04377104 -1.18148148 -0.90606061 -0.90606061]
```

Tutorial Question 3 – Supervised

Question 3

What is a difference, between a supervised, semi-supervised, and unsupervised anomaly detection techniques?

Supervised

- With labels, i.e., normal objects vs. outliers
- Binary classification problem
- Extremely imbalanced, majority of the objects are normal.

Unsupervised

- Labels are not available
- Clustering based methods

Semi-supervised

- Some labeled examples is feasible, the number of such labeled examples is small.
- Combining classification-based and clustering based methods
- Use the one-class model to label normal objects, e.g. One-class SVM