Anomaly Detection

Density Based Methods

Paul Irofti
Andrei Pătrașcu
Cristian Rusu

Computer Science Department
Faculty of Mathematics and Computer Science
University of Bucharest
Email: `first.last@fmi.unibuc.ro`

Outline

- clustering, distance and density
- k-Nearest Neighbors
- Local Outlier Factor
- Local Correlation Integral

The course references are Aggarwal 2017, Ch.4 with initial papers for K-NN by Cover and Hart 1967 and LOF by Breunig et al. 2000.

Algorithm Types

Given a data point, discriminate based on

- Clustering
  - non-membership to any data cluster of the data point
  - distance to other clusters
  - size of the closest cluster
  - binary: either belongs to cluster else is an anomaly
- Distance
  - proximity: distance to its k-nearest neighbor (KNN)
  - variants change distance type or average the distance score
  - large KNN distances define the anomalies
  - high granularity results
  - high algorithmic complexity (e.g. $O(N^2)$)
- Density
  - split data space into regions
  - compute the local density of each region
  - data density is turned into anomaly score for each point
  - clustering partitions data-points, density partitions data-space

Data Points vs Data Space

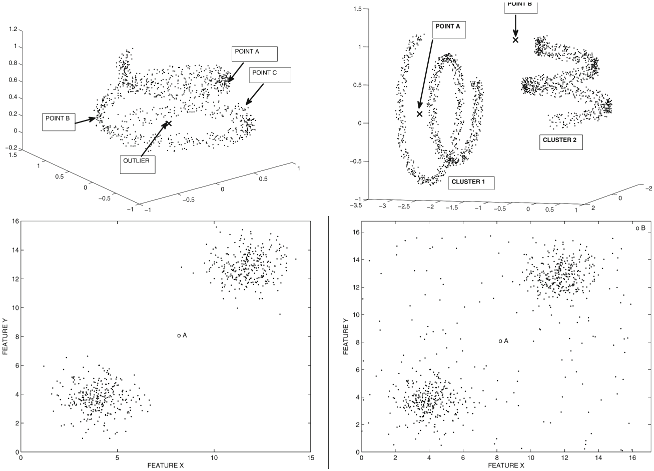


Figure: Data points and data space (Aggarwal 2017)

k-Nearest Neigbors

Nearest Neighbors

Definition

Exact KNN. The anomaly score of a point x is given by its distance to its k-th nearest neighbor.

Assumption: anomalous data points are further away than normal data points.

Example

We can identify small isolated clusters of k_0 anomalous data-points by selecting a value $k \geq k_0$ in the KNN algorithm.

KNN: the choice of k_0

Example

We can identify small isolated clusters of k_0 anomalous data-points by selecting a value $k \geq k_0$ in the KNN algorithm.

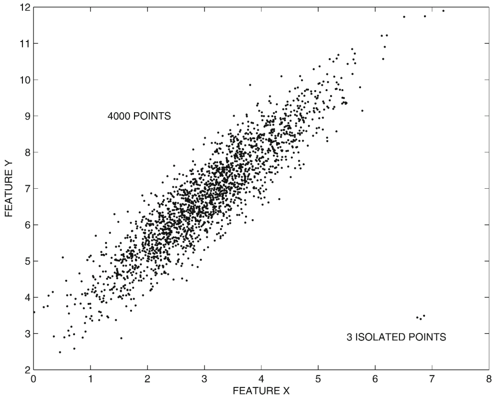


Figure: The choice of $k \geq k_0$ (Aggarwal 2017)

Nearest Neighbors Algorithm

How do we implement this?

Nearest Neighbors Algorithm

How do we implement this?

1. Choose a data-point:

Nearest Neighbors Algorithm

How do we implement this?

1. Choose a data-point: Let x be an m-dimensional point from the dataset $X \in \mathbb{R}^{m \times N}$.

2. Choose a distance function:

Nearest Neighbors Algorithm

How do we implement this?

1. Choose a data-point: Let x be an m-dimensional point from the dataset $X \in \mathbb{R}^{m \times N}$.

2. Choose a distance function: $\|x - y\|_2$ where $y \neq x$ and $y \in X$.

3. Compute the distances:

Nearest Neighbors Algorithm

How do we implement this?

1. Choose a data-point: Let x be an m-dimensional point from the dataset $X \in \mathbb{R}^{m \times N}$.

2. Choose a distance function: $\|x - y\|_2$ where $y \neq x$ and $y \in X$.

3. Compute the distances:
$dist(x) = \{s \mid s = \|x - y\|_2\,,\ \forall y \in X,\ y \neq x\}$.

4. Set the anomaly score:

Nearest Neighbors Algorithm

How do we implement this?

1. Choose a data-point: Let x be an m-dimensional point from the dataset $X \in \mathbb{R}^{m \times N}$.

2. Choose a distance function: $\|x - y\|_2$ where $y \neq x$ and $y \in X$.

3. Compute the distances:
$dist(x) = \{s \mid s = \|x - y\|_2\,,\ \forall y \in X,\ y \neq x\}$.

4. Set the anomaly score: $knn(x) = min_k(dist(x))$ where $min_k(\cdot)$ is the function computing the k-th smallest number in a set.

5. Repeat steps 1–4 for all points in X.

KNN Granularity

What is the complexity of the above?

KNN Granularity

What is the complexity of the above? $O(N^2)$!

KNN Granularity

What is the complexity of the above? $O(N^2)$!

Remark

Granularity. Distance-based methods have a higher granularity compared to cluster based methods. We compare each point to the rest of the points, whereas for clusters we only compare with the centroids.

KNN Granularity

What is the complexity of the above? $O(N^2)$!

Remark

Granularity. Distance-based methods have a higher granularity compared to cluster based methods. We compare each point to the rest of the points, whereas for clusters we only compare with the centroids.

Mitigations:

- pre-select a sample of data points $\tilde{N} \ll N$
- all N points are scored based on these $\tilde{N}$ scores
- smoothing or averaging techniques can be applied post-processing to reduce sensibility to choice of $\tilde{N}$
- converges to a sort of clustering method

KNN Parameters

How do we choose k?

KNN Parameters

How do we choose k?

Kind reminder. In anomaly detection we are in the unsupervised setting where we can not perform parameter tuning.

KNN Parameters

How do we choose k?

Kind reminder. In anomaly detection we are in the unsupervised setting where we can not perform parameter tuning.

General approach. Naturally we train multiple KNN models where we vary k and create an ensemble where voting methods are employed in order to decide which data is anomalous.

KNN Parameters

How do we choose k?

Kind reminder. In anomaly detection we are in the unsupervised setting where we can not perform parameter tuning.

General approach. Naturally we train multiple KNN models where we vary k and create an ensemble where voting methods are employed in order to decide which data is anomalous.

Particular approach. Smoothen the anomaly score so that it is less sensible to a particular choice of k.

KNN Variants: Average KNN Scores

Definition

Average KNN score. The anomaly score of $x \in X$ is its average distance to its k-nearest neighbors.

Average KNN is:

- better suited for unsupervised grid-search where a range of k's are used
- it is less sensitive to the particular choices for k
- averages Exact-KNN over a range of k
- provides worse results than the true k value in the Exact-KNN variant

Formally, $avgknn(x) = \mu_k(dist(X))$, where $\mu_k(\cdot)$ is the average of the smallest k numbers in the set.

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

Harmonic KNN is:

- harmonic mean is very sensible to small distances. why?

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

Harmonic KNN is:

- harmonic mean is very sensible to small distances. why?
- we can use greater values of k. how can we profit?

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

Harmonic KNN is:

- harmonic mean is very sensible to small distances. why?
- we can use greater values of k. how can we profit?
- use $k = N$ and be parameter-free!

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

Harmonic KNN is:

- harmonic mean is very sensible to small distances. why?
- we can use greater values of k. how can we profit?
- use $k = N$ and be parameter-free!
- provides good results

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

Harmonic KNN is:

- harmonic mean is very sensible to small distances. why?
- we can use greater values of k. how can we profit?
- use $k = N$ and be parameter-free!
- provides good results
- it is no longer sensitive to the particular choices for k

KNN Variants: Harmonic KNN Scores

Definition

Harmonic KNN score. The anomaly score of $x \in X$ is its harmonic mean of its k-nearest neighbors distances.

The harmonic mean is $H(X) = \frac{n}{\sum_i \frac{1}{x_i}}$.

Formally, $hknn(x) = H_k(dist(X))$, where $H_k(\cdot)$ is the harmonic mean of the smallest k numbers in the set.

Harmonic KNN is:

- harmonic mean is very sensible to small distances. why?
- we can use greater values of k. how can we profit?
- use $k = N$ and be parameter-free!
- provides good results
- it is no longer sensitive to the particular choices for k
- worse results than the true k value in the Exact-KNN

Local Outlier Factor

Problems: Density and Cluster Orientation

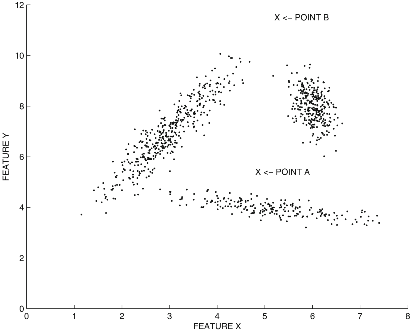


Figure: Data Locality (Aggarwal 2017)

- impacts data density and cluster orientation
- varying density across data space
- distance-based limitations when density variation is high

Example: Distance versus Locality

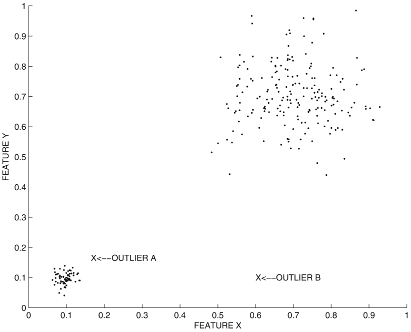


Figure: Distance versus Locality (Aggarwal 2017)

- two cluster with different sparsity
- A requires small distance threshold
- if k is small, then lots of false-positives in the sparse cluster
- need multiple distance thresholds in heterogeneous data distributions

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$:

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$: when y inside dense region and x is far away

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \le knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$: when y inside dense region and x is far away
- $knn(y)$:

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$: when y inside dense region and x is far away
- $knn(y)$: when x and y are close, but $knn(y)$ is large; smoothing!

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$: when y inside dense region and x is far away
- $knn(y)$: when x and y are close, but $knn(y)$ is large; smoothing!
- larger values of k will bring a greater smoothing effect

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$: when y inside dense region and x is far away
- $knn(y)$: when x and y are close, but $knn(y)$ is large; smoothing!
- larger values of k will bring a greater smoothing effect
- reachability distances become similar with large k's

Local Outlier Factor (LOF)

Let $L_k(\cdot)$ be the set of points that are the knn of a given point:

$$L_k(x) = \{y \mid \|x - y\| \leq knn(x),\ \forall y \in X\}$$

this can contain more than k points, but not less!

Let $R_k(\cdot, \cdot)$ be the reachability distance of two points:

$$R_k(x, y) = \max\{\|x - y\|, knn(y)\}$$

- $\|x - y\|$: when y inside dense region and x is far away
- $knn(y)$: when x and y are close, but $knn(y)$ is large; smoothing!
- larger values of k will bring a greater smoothing effect
- reachability distances become similar with large k's
- reachability is not symmetric!

LOF: Reachability

Reachability is not symmetric!

LOF: Reachability

Reachability is not symmetric!

Proof in class

LOF: Average Reachability Distance

We now define the average reachability of a point in regards to its KNN's:

$$AR_k(x) = \mu_{y \in L_k(x)} R_k(x, y)$$

where μ is the average of each pair $R_k(x, y)$ with $y \in L_k(x)$.

The inverse of AR_k is defined as the reachability density.

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

- the sum of distance ratios act as data normalization

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

- the sum of distance ratios act as data normalization
- LOF score is normalized reachability distance of a given point

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

- the sum of distance ratios act as data normalization
- LOF score is normalized reachability distance of a given point
- normalization factor is the harmonic mean

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

- the sum of distance ratios act as data normalization
- LOF score is normalized reachability distance of a given point
- normalization factor is the harmonic mean
- homogeneous distributions have $LOF \approx 1$

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

- the sum of distance ratios act as data normalization
- LOF score is normalized reachability distance of a given point
- normalization factor is the harmonic mean
- homogeneous distributions have $LOF \approx 1$
- this solves the problem in Figure 4

LOF: Scoring

The local outlier factor is the mean average reachability of x compared to its neighbors average:

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = AR_k(x) \mu_{y \in L_k(x)} \frac{1}{AR_k(y)}$$

- the sum of distance ratios act as data normalization
- LOF score is normalized reachability distance of a given point
- normalization factor is the harmonic mean
- homogeneous distributions have $LOF \approx 1$
- this solves the problem in Figure 4
- anomalous points have $LOF \gg 1$

Example: LOF Scoring

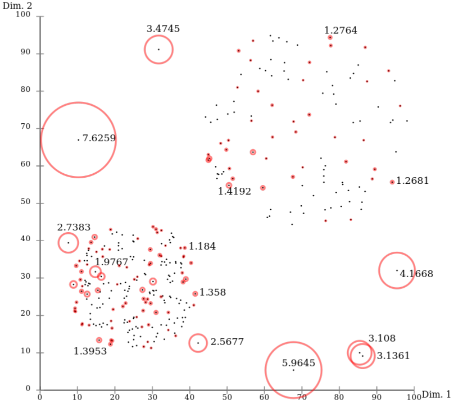
LOF normalization factor is the harmonic mean

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = \frac{AR_k(x)}{H_{y \in L_k(x)} AR_k(y)}$$

Example: LOF Scoring

LOF normalization factor is the harmonic mean

$$LOF_k(x) = \mu_{y \in L_k(x)} \frac{AR_k(x)}{AR_k(y)} = \frac{AR_k(x)}{H_{y \in L_k(x)} AR_k(y)}$$



Source: https://en.wikipedia.org/wiki/Local_outlier_factor

LOF: Properties

A few remarks about LOF:

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances
- pseudo-density: inverse of smoothed reachability

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances
- pseudo-density: inverse of smoothed reachability
- its a relaxed version of density (see LOCI method)

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances
- pseudo-density: inverse of smoothed reachability
- its a relaxed version of density (see LOCI method)

Variants

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances
- pseudo-density: inverse of smoothed reachability
- its a relaxed version of density (see LOCI method)

Variants

- raw distances instead of reachability distances

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances
- pseudo-density: inverse of smoothed reachability
- its a relaxed version of density (see LOCI method)

Variants

- raw distances instead of reachability distances
- arithmetic mean instead of harmonic mean

LOF: Properties

A few remarks about LOF:

- we can use other means of smoothing for normalization
- can be seen as relative distance-based approach with smoothing
- adjusts to varying data density using relative distances
- pseudo-density: inverse of smoothed reachability
- its a relaxed version of density (see LOCI method)

Variants

- raw distances instead of reachability distances
- arithmetic mean instead of harmonic mean
- local distance-based outlier factor (LDOF) by Zhang, Hutter, and Jin 2009 uses averaged pairwise distances from L_k

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!
- solution: use other metrics than harmonic mean

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!
- solution: use other metrics than harmonic mean
- solution: use regularization

$$LOF_k(x) = \frac{\alpha + AR_k(x)}{\alpha + H_{y \in L_k(x)} AR_k(y)}$$

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!
- solution: use other metrics than harmonic mean
- solution: use regularization

$$LOF_k(x) = \frac{\alpha + AR_k(x)}{\alpha + H_{y \in L_k(x)} AR_k(y)}$$

Choosing k:

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!
- solution: use other metrics than harmonic mean
- solution: use regularization

$$LOF_k(x) = \frac{\alpha + AR_k(x)}{\alpha + H_{y \in L_k(x)} AR_k(y)}$$

Choosing k:

- Breunig et al. 2000 recommend choosing $\max L_k(x)$ after grid-search

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!
- solution: use other metrics than harmonic mean
- solution: use regularization

$$LOF_k(x) = \frac{\alpha + AR_k(x)}{\alpha + H_{y \in L_k(x)} AR_k(y)}$$

Choosing k:

- Breunig et al. 2000 recommend choosing $\max L_k(x)$ after grid-search
- tightly coupled data within a single data distribution will impact the scores

LOF: Data and Parameters

Duplicate points in X:

- duplicate data in X will result in zero harmonic mean!
- solution: use other metrics than harmonic mean
- solution: use regularization

$$LOF_k(x) = \frac{\alpha + AR_k(x)}{\alpha + H_{y \in L_k(x)} AR_k(y)}$$

Choosing k:

- Breunig et al. 2000 recommend choosing $\max L_k(x)$ after grid-search
- tightly coupled data within a single data distribution will impact the scores
- small values of k increases false-positive risks

Local Correlation Integral (LOCI)

LOCI (Papadimitriou et al. 2003) is a close relative of LOF that uses standard density.

Local Correlation Integral (LOCI)

LOCI (Papadimitriou et al. 2003) is a close relative of LOF that uses standard density.

Let the counting neighborhood of data point x be:

$$M(x, \varepsilon) = |\{y \mid \|x - y\| \leq \varepsilon, \ \forall y \in X\}|$$

where $| \cdot |$ is the sets cardinality operator.

Local Correlation Integral (LOCI)

LOCI (Papadimitriou et al. 2003) is a close relative of LOF that uses standard density.

Let the counting neighborhood of data point x be:

$$M(x, \varepsilon) = |\{y \mid \|x - y\| \leq \varepsilon,\ \forall y \in X\}|$$

where $|\cdot|$ is the sets cardinality operator.

Then the average density in the δ-neighborhood of x

$$AM(x, \varepsilon, \delta) = \mu_{y:\|x-y\| \leq \delta} M(y, \varepsilon)$$

where $\delta > \varepsilon$ is the sampling neighborhood of x.

Local Correlation Integral (LOCI)

LOCI (Papadimitriou et al. 2003) is a close relative of LOF that uses standard density.

Let the counting neighborhood of data point x be:

$$M(x, \varepsilon) = |\{y \mid \|x - y\| \leq \varepsilon,\ \forall y \in X\}|$$

where $|\cdot|$ is the sets cardinality operator.

Then the average density in the δ-neighborhood of x

$$AM(x, \varepsilon, \delta) = \mu_{y:\|x-y\|\leq\delta} M(y, \varepsilon)$$

where $\delta > \varepsilon$ is the sampling neighborhood of x.

In practice we choose $\varepsilon = c\delta$ where $c = \frac{1}{2}$ is a popular choice.

LOCI: Scoring

Let us now define the equivalent neighborhood-aware averaging score of a point

$$MDEF(x, \varepsilon, \delta) = 1 - \frac{M(x, \varepsilon)}{AM(X, \varepsilon, \delta)}$$

called *multi-granularity deviation factor*.

LOCI: Scoring

Let us now define the equivalent neighborhood-aware averaging score of a point

$$MDEF(x, \varepsilon, \delta) = 1 - \frac{M(x, \varepsilon)}{AM(X, \varepsilon, \delta)}$$

called *multi-granularity deviation factor*.

The greater the MDEF value, the greater its anomalous score.

LOCI: Scoring

Let us now define the equivalent neighborhood-aware averaging score of a point

$$MDEF(x, \varepsilon, \delta) = 1 - \frac{M(x, \varepsilon)}{AM(X, \varepsilon, \delta)}$$

called *multi-granularity deviation factor*.

The greater the MDEF value, the greater its anomalous score.

Binary labels are obtained through the use of standard deviation

$$\sigma(x, \varepsilon, \delta) = \frac{STD_{y:\|x-y\| \leq \delta} M(x, \varepsilon)}{AM(X, \varepsilon, \delta)}$$

where *STD* computes the standard deviation of the sampling neighborhood.

LOCI: Scoring

Let us now define the equivalent neighborhood-aware averaging score of a point

$$MDEF(x, \varepsilon, \delta) = 1 - \frac{M(x, \varepsilon)}{AM(X, \varepsilon, \delta)}$$

called *multi-granularity deviation factor*.

The greater the MDEF value, the greater its anomalous score.

Binary labels are obtained through the use of standard deviation

$$\sigma(x, \varepsilon, \delta) = \frac{STD_{y:\|x-y\| \le \delta} M(x, \varepsilon)}{AM(X, \varepsilon, \delta)}$$

where STD computes the standard deviation of the sampling neighborhood.

In practice $MDEF \geq k\sigma$ is used with $k = 3$ being a popular choice inspired from statistics.

LOCI: Parameters

Choosing ε and δ:

LOCI: Parameters

Choosing ε and δ:

- only one parameter if $\varepsilon = c\delta$ with $c = \frac{1}{2}$

LOCI: Parameters

Choosing ε and δ:

- only one parameter if $\varepsilon = c\delta$ with $c = \frac{1}{2}$
- grid-search multiple values of δ

LOCI: Parameters

Choosing ε and δ:

- only one parameter if $\varepsilon = c\delta$ with $c = \frac{1}{2}$
- grid-search multiple values of δ
- popular options is to start with neighborhoods of 20 up to N

LOCI: Parameters

Choosing ε and δ:

- only one parameter if $\varepsilon = c\delta$ with $c = \frac{1}{2}$
- grid-search multiple values of δ
- popular options is to start with neighborhoods of 20 up to N
- anomaly if MDEF is large within **any** of the δ settings

LOCI: Parameters

Choosing ε and δ:

- only one parameter if $\varepsilon = c\delta$ with $c = \frac{1}{2}$
- grid-search multiple values of δ
- popular options is to start with neighborhoods of 20 up to N
- anomaly if MDEF is large within **any** of the δ settings
- sub-sample considered neighborhoods based on invariance to δ choice

References

Aggarwal, Charu C (2017). *An introduction to outlier analysis*. Springer.

Breunig, Markus M et al. (2000). "LOF: identifying density-based local outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

Cover, Thomas and Peter Hart (1967). "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1, pp. 21–27.

Papadimitriou, Spiros et al. (2003). "Loci: Fast outlier detection using the local correlation integral". In: *Proceedings 19th international conference on data engineering (Cat. No. 03CH37405)*. IEEE, pp. 315–326.

Zhang, Ke, Marcus Hutter, and Huidong Jin (2009). "A new local distance-based outlier detection approach for scattered real-world data". In: *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13*. Springer, pp. 813–822.