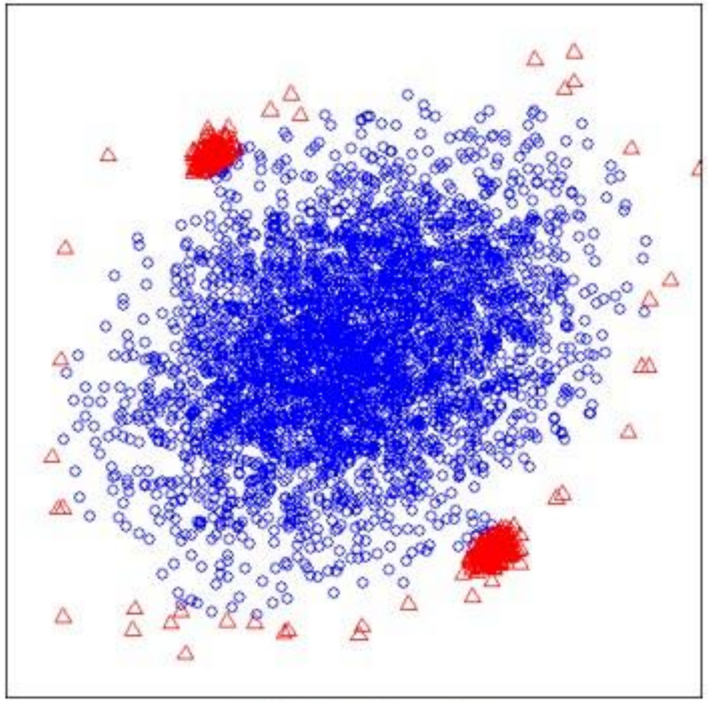
Introduction to Anomaly Detection

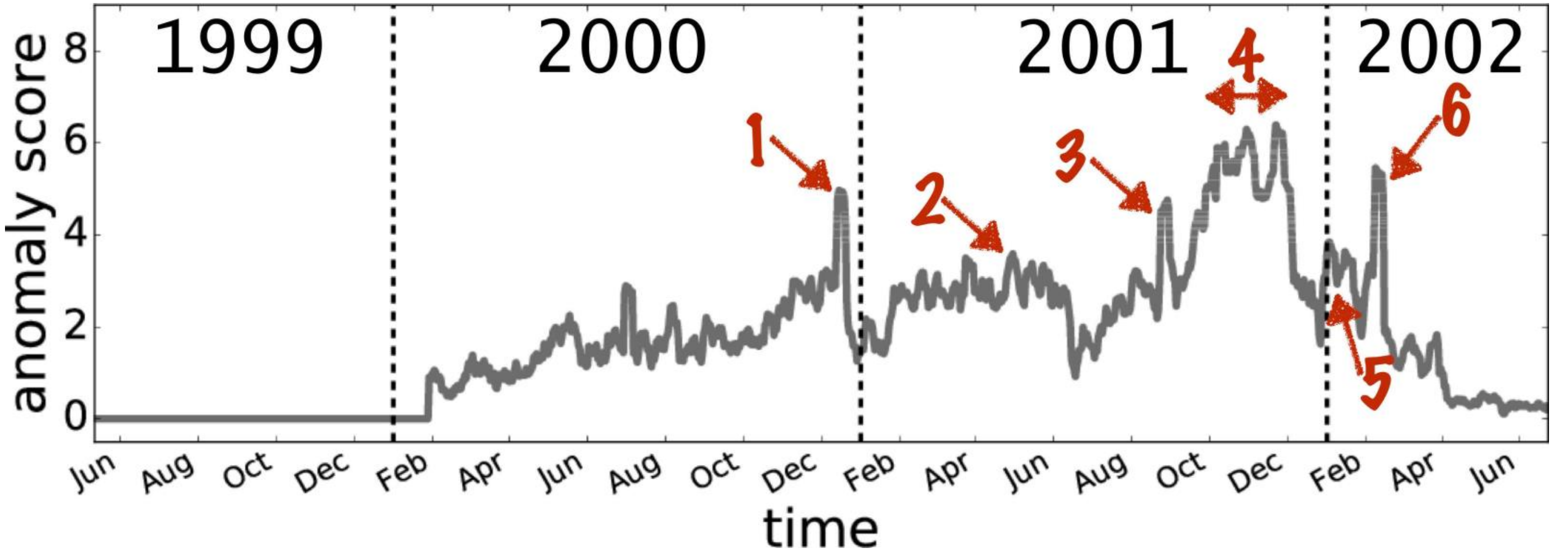
Linghao Chen

HOMEPAGE: https://lhchen.top
lhchen@stu.xidian.edu.cn
School of Computer Science and Technology, Xidian University, Xi'an, ShaanXi, P.R.China

Anomaly Detection

What is it?





[1]: Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." *2008 eighth ieee international conference on data mining*. IEEE, 2008.
[2]: Eswaran, Dhivya, et al. "Spotlight: Detecting anomalies in streaming graphs." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

XIDIAN UNIVERSITY

Problems

Why so hard to detect anomaly?

- ✓ Unsupervised learning in most cases;
- ✓ The data is extremely unbalanced;
- ✓ It often involves density estimation, which requires a large amount of distance or similarity calculations, and computationally expensive;
- ✓ Real-time detection;
- ✓ Interpretability of methods.

Methods

Classic Methods:

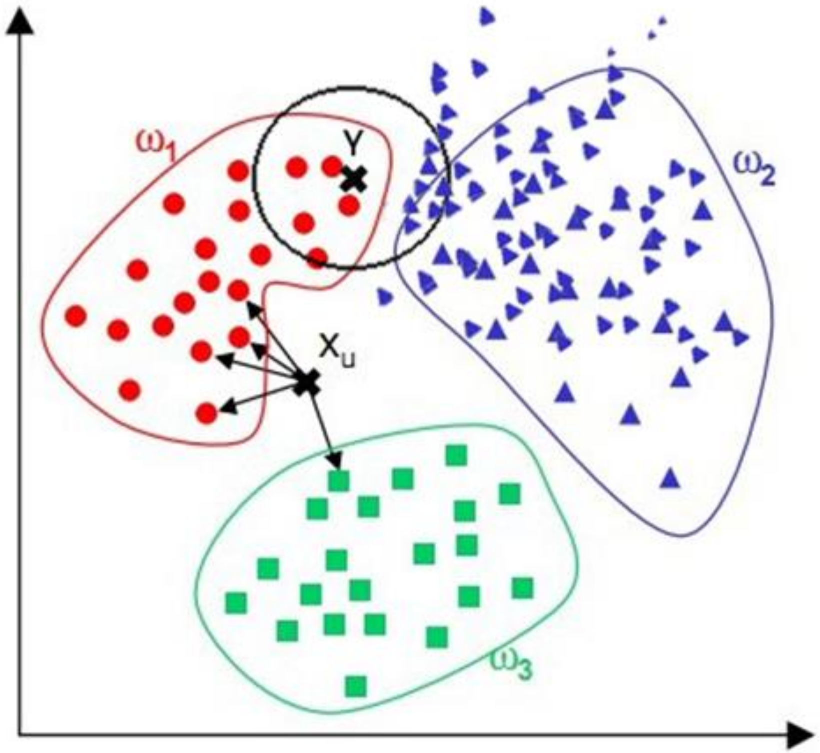
- kNN(K-Nearest Neighbor)
- LOF(Local Outlier Factor)
- PCA(Principal Component Analysis)
- HBOS(Histogram-based Outlier Score)
- Isolation Forest
- AE(Auto Encoder)

kNN(K-Nearest Neighbor)

$$Dis(x,y) = \left(\sum_{i=1}^{N} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Choose Top K-th Dstance

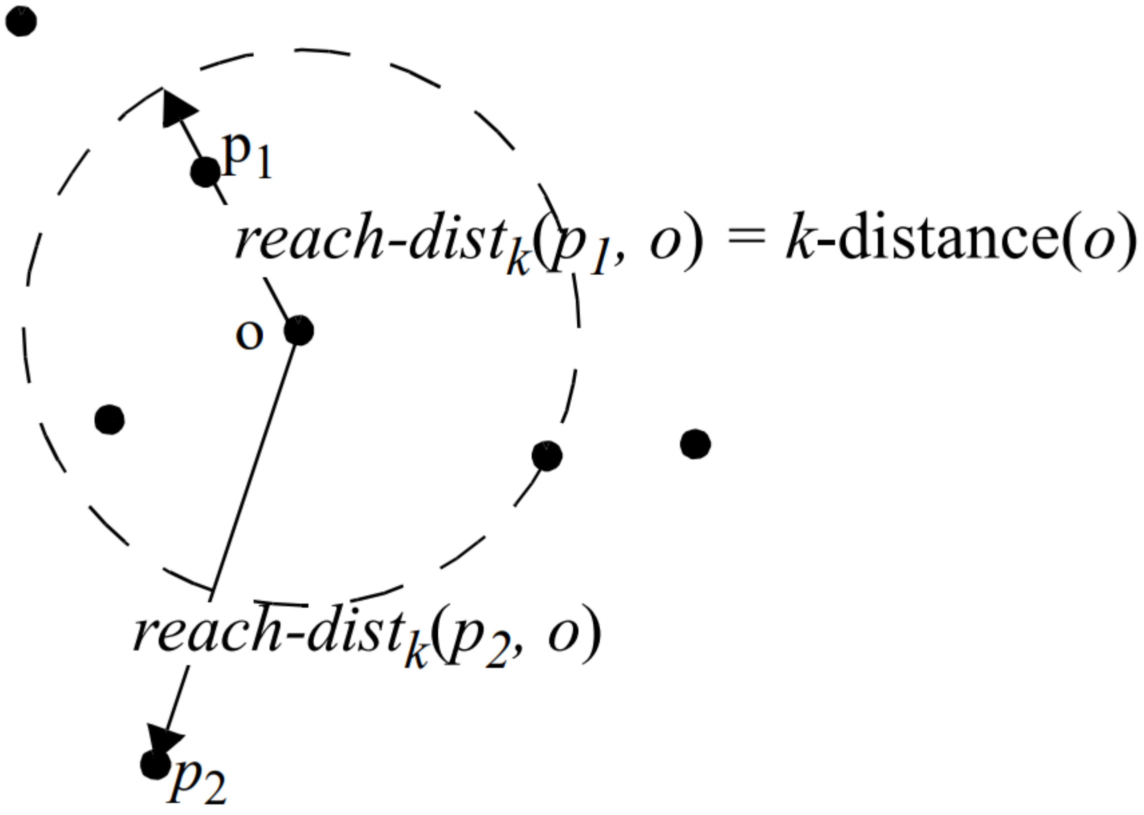
Simple but expensive!

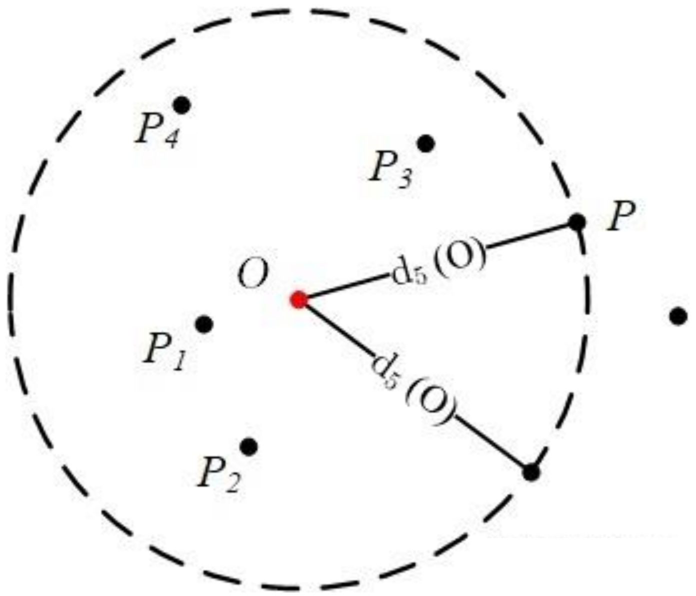


[1]: Ramaswamy, S., Rastogi, R. and Shim, K., 2000, May. Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record, 29(2), pp. 427-438.

LOF(Local Outlier Factor)

K-distance of an object p





5-distance

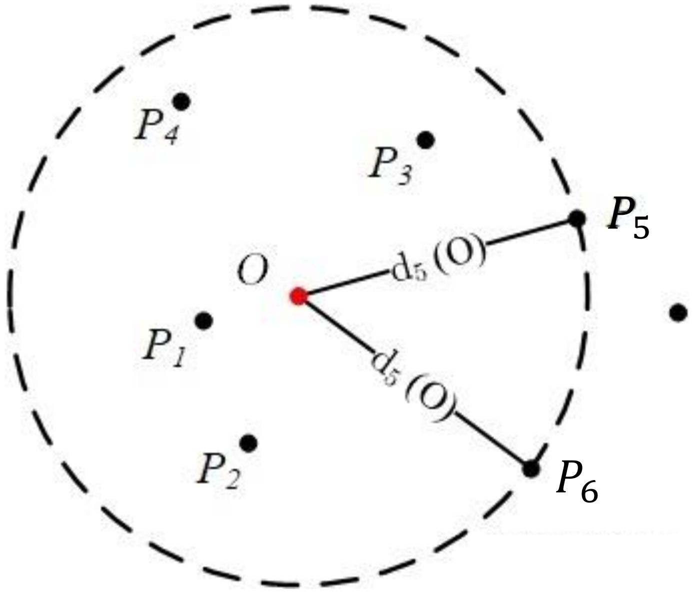
[1]: Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000, May. LOF: identifying density-based local outliers. ACM Sigmod Record, 29(2), pp. 93-104.

LOF(Local Outlier Factor)

K-distance neighborhood of an object p

$$N_k(O) = \{P' \in D\{O\} \mid d(O,P') \leq d_k(O)\}$$

$$N_5(O) = \{P_1, P_2, P_3, P_4, P_5, P_6\}$$



5-distance

Reachability distance of an object P w.r.t. object O

$$\rho_k(P) = \frac{|N_k(P)|}{\sum_{O \in N_k(P)} d\_k(P,O)}$$

$$LOF_k(P) = \frac{\sum_{O \in N_k(P)} \frac{\rho_k(O)}{\rho_k(P)}}{|N_k(P)|}$$

[1]: Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000, May. LOF: identifying density-based local outliers. ACM Sigmod Record, 29(2), pp. 93-104.

PCA(Principal Component Analysis)

Algorithm

Input: $X \in \mathbb{R}_{\mathrm{n} \times m}$ with n samples

Output: $Y = WX \in \mathbb{R}_{\mathrm{n} \times m'}$

Normalization: $x_i = x_i - \frac{1}{m} \sum_{j=1}^{m} x_j$

Covariance matrix: $C = \frac{1}{m} X X^T$

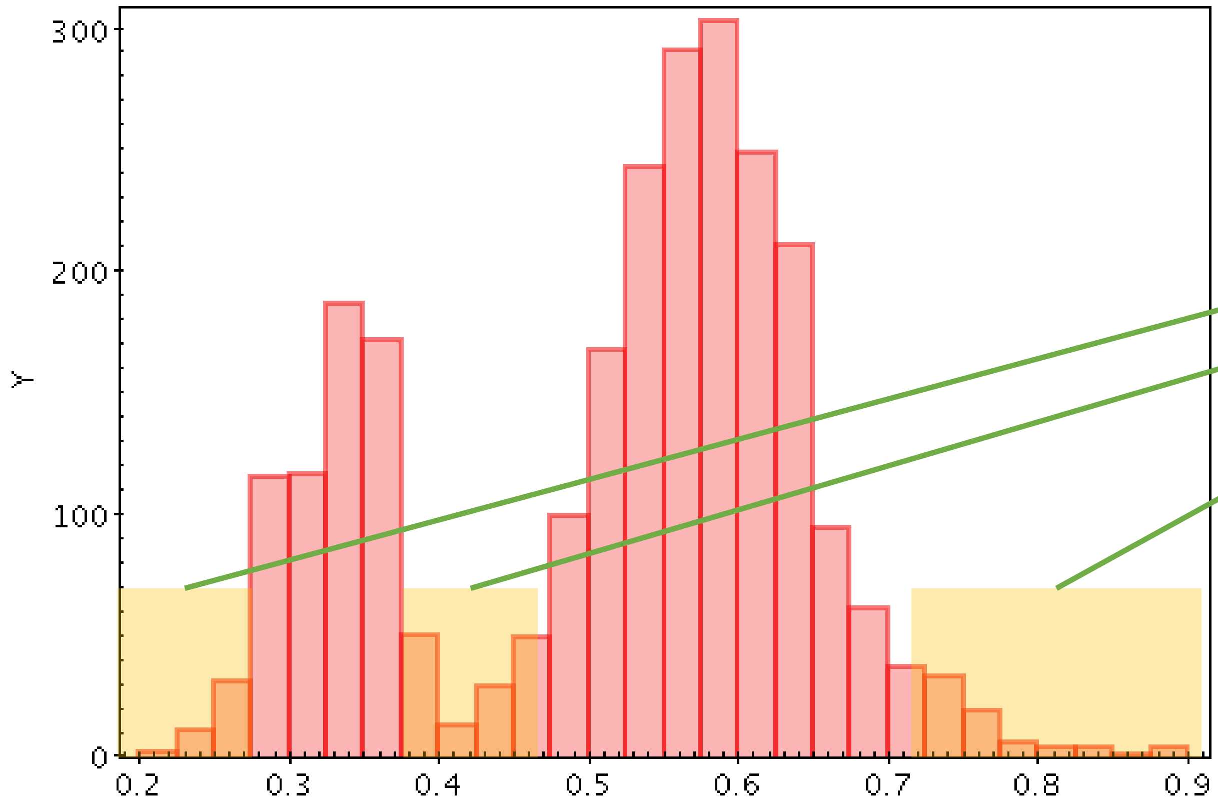
Calculate eigenvectors

Anomaly score: the distance between the abnormal sample and the feature vector

[1]: Shyu, Mei-Ling, et al. A novel anomaly detection scheme based on principal component classifier. MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2003.

HBOS(Histogram-based Outlier Score)

Methods



Low density area

[1]: Goldstein, M. and Dengel, A., 2012. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In KI-2012: Poster and Demo Track, pp.59-63.

HBOS(Histogram-based Outlier Score)

Assumption

Multidimensional data is ***independent*** of each dimension.

Algorithm

- Draw a data histogram
- Divide the value range into ***K*** buckets of equal(sometimes can be dynamic) width, and the frequency of the value falling into each bucket is used as an estimate of density.

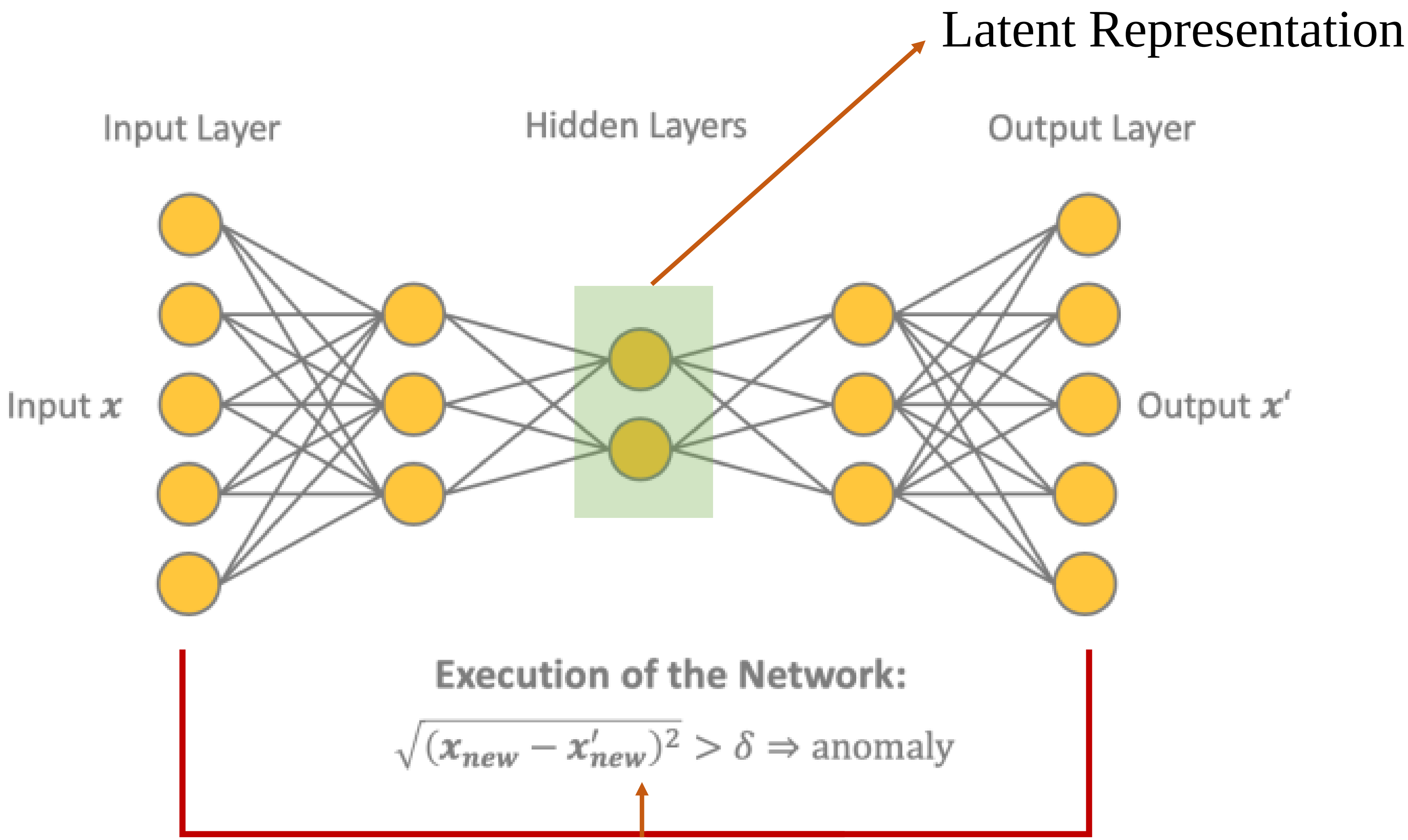
Anomaly Score

$$HBOS(p) = \sum_{i=0}^{a} \log(\frac{1}{hist_i(p)})$$

[1]: Goldstein, M. and Dengel, A., 2012. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In KI-2012: Poster and Demo Track, pp.59-63.

西安電子科技大學 XIDIAN UNIVERSITY

AE(Auto Encoder)



[1]: Ramaswamy, S., Rastogi, R. and Shim, K., 2000, May. Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record, 29(2), pp. 427-438.

Isolation Forest

Isolation Forest

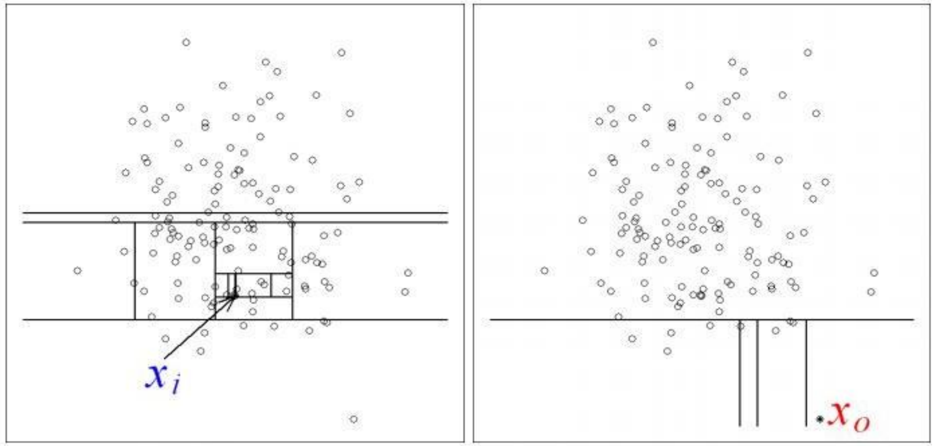
Fei Tony Liu, Kai Ming Ting
Gippsland School of Information Technology
Monash University, Victoria, Australia
{tony.liu},{kaiming.ting}@infotech.monash.edu.au

Zhi-Hua Zhou
National Key Laboratory
for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@lamda.nju.edu.cn

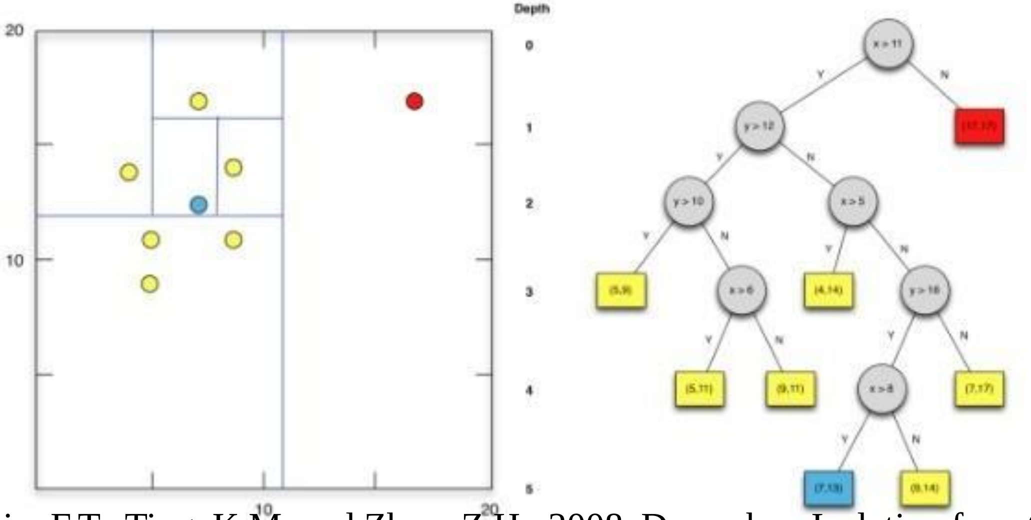
ICDM '08

[1]: Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In International Conference on Data Mining (ICDM), pp. 413-422. IEEE

Anomaly Detection



(a) Isolating x_i (b) Isolating x_o



Algorithm 2 : $iTree(X, e, l)$

Inputs: X - input data, e - current tree height, l - height limit

Output: an iTree

if $e \geq l$ or $|X| \leq 1$ **then**
  return *exNode*$\{Size \leftarrow |X|\}$
else
  let Q be a list of attributes in X
  randomly select an attribute $q \in Q$
  randomly select a split point p from max and min values of attribute q in X
  $X_l \leftarrow filter(X, q < p)$
  $X_r \leftarrow filter(X, q \geq p)$
  return *inNode*$\{Left \leftarrow iTree(X_l, e+1, l),$
    $Right \leftarrow iTree(X_r, e+1, l),$
    $SplitAtt \leftarrow q,$
    $SplitValue \leftarrow p\}$
end if

---

[1]: Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In International Conference on Data Mining (ICDM), pp. 413-422. IEEE

Anomaly Detection

Algorithm 1 : $iForest(X, t, \psi)$

Inputs: X - input data, t - number of trees, ψ - sub-sampling size

Output: a set of t $iTrees$

1: **Initialize** $Forest$
2: set height limit $l = ceiling(\log_2 \psi)$
3: **for** $i = 1$ to t **do**
4: $\quad X' \leftarrow sample(X, \psi)$
5: $\quad Forest \leftarrow Forest \cup iTree(X', 0, l)$
6: **end for**
7: **return** $Forest$

[1]: Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In International Conference on Data Mining (ICDM), pp. 413-422. IEEE

Isolation Forest

Anomaly Detection

Algorithm 3 : $PathLength(x, T, e)$

Inputs : x - an instance, T - an iTree, e - current path length; to be initialized to zero when first called

Output: path length of x

```
if T is an external node then
    return e + c(T.size) {c(.) is defined in Equation 1}
end if
a ← T.splitAtt
if x_a < T.splitValue then
    return PathLength(x, T.left, e + 1)
else {x_a ≥ T.splitValue}
    return PathLength(x, T.right, e + 1)
end if
```

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

$$c(n) = 2H(n-1) - (2(n-1)/n), \qquad (1)$$

where $H(i)$ is the harmonic number and it can be estimated by $\ln(i)$ + 0.5772156649 (Euler's constant). As $c(n)$ is the average of $h(x)$ given n, we use it to normalise $h(x)$. The anomaly score s of an instance x is defined as:

[1]: Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In International Conference on Data Mining (ICDM), pp. 413-422. IEEE

REFERENCE

[1]: Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." 2008 eighth ieee international conference on data mining. IEEE, 2008.

[2]: Eswaran, Dhivya, et al. "Spotlight: Detecting anomalies in streaming graphs." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

[3]: Ramaswamy, S., Rastogi, R. and Shim, K., 2000, May. Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record, 29(2), pp. 427-438.

[4]: Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000, May. LOF: identifying density-based local outliers. ACM Sigmod Record, 29(2), pp. 93-104.

[5]: Shyu, Mei-Ling, et al. A novel anomaly detection scheme based on principal component classifier. MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2003.

[6]: Goldstein, M. and Dengel, A., 2012. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In KI-2012: Poster and Demo Track, pp.59-63.

[7]: Ramaswamy, S., Rastogi, R. and Shim, K., 2000, May. Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record, 29(2), pp. 427-438.

[8]: Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In International Conference on Data Mining (ICDM), pp. 413-422. IEEE

Q&A

Linghao Chen
HOMEPAGE: https://lhchen.top
lhchen@stu.xidian.edu.cn
School of Computer Science and Technology, Xidian University, Xi'an, ShaanXi, P.R.China

西安电子科技大学
XIDIAN UNIVERSITY