Institut de Mathématiques de Toulouse, INSA Toulouse

Anomaly Detection

ML Training- CERFACS
October, 2021

Béatrice Laurent

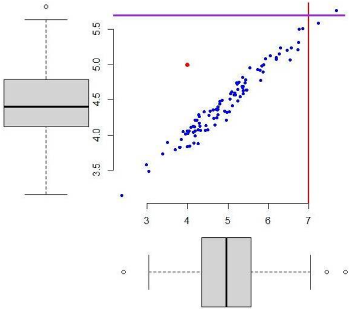
Outline

- Introduction
- Different aspects of an anomaly detection problem
- Focus on some anomaly detection algorithms
  - Isolation Forest
  - Local Outlier Factor
  - One Class SVM
- Conclusion

Introduction

- Anomaly detection refers to the problem of finding patterns in data that do not conform the expected normal behavior.
- These anomalous patters are referred as anomalies, outliers, defects, novelty ...
- Anomalies (or *outliers*) detection is a major issue for many application fields, in particular for industrial applications of statistics : failure detection, manufacturing defects ideally for predictive maintenance (failure prediction before it happens) ..
- Outliers detection also finds many applications in the following domains : credit card, insurance, tax fraud detection, intrusion detection for cyber security.
- Detection of abnormal behaviors on sensors (functional data) is also a major issue for a wide range of applications in aeronautics and space, health monitoring or monitoring elderly people for example.

Outliers refer to patterns in the data that do not conform to a well defined notion of *normal behavior*. Here is an example in dimension 2, where outliers detected simultaneously in one dimensional boxplots seem less anomalous than the red point in view of the two-dimensional repartition of the data.



- Anomalies are present in data sets for various reasons, but the common point is that these reasons present an interest for the analyst.
- This should be distinguished from *noise removal*. Noise in the data will degrade the performances of statistical inference, it has no significance by itself and should be removed before or during the data analysis.
- Another related topic is the *novelty detection* which aims at finding unseen patterns in the data. This concerns mostly temporal or spatial data.

Challenges

Anomaly detection is most of the time very challenging for various reasons :

- Defining a *normal region* containing every possible normal behavior is generally difficult since
  - The boundary between normal and outlying behavior is generally fuzzy
  - Normal behavior may evolve with time
- The notion of outlier depends on the application domain
- Labeled data for training/validation are rarely available
- When the outliers result from malicious actions, the adversary adapts themselves to confound the outlier with to the normal behavior
- Often, the data contain noise, which may be difficult to distinguish from the anomalies.

Challenges

- This explains why most anomaly detection techniques focus on a specific formulation depending on the nature of the data, the application domain, the type of outliers to be detected, the representation of the normal behavior ..
- The literature on this subject is very abundant. We do not have the pretension in this course to cover all the possible cases, but to present a focus on some of the most "classical" methods.

Outline

- Introduction
- Different aspects of an anomaly detection problem
- Focus on some anomaly detection algorithms
- Conclusion

The input data

The nature of the input data is an important aspect of anomaly detection. We can have various types of data

- Univariate or more generally multidimensional variables that may be of various types (binary, categorial or continuous)
- Time series or functional data
- Spatial data
- Images
- Genome sequences ..

An important point is that the features used by an outlier detection technique are not necessarily the observable variables : several methods use a preprocessing on the input data to extract features, ideally which are more likely to discriminate between the normal and the outlying behavior in the data (such as wavelets transform on functional data or images for example).

Labels

In addition with the input data, an anomaly detection algorithm may also have additional informations, such as a labelled training data set, specifiying if each observation is *normal or anomalous*. Labeling is often done manually by human experts, hence labeled data are often not available. Depending on the type of observations it will use, an anomaly detection algorithm fall into three categories :

- Supervised anomaly detection : We assume here that labeled data for both classes -normal and anomalous-are available. In this case, we will be able to built predictive models.
- Semi supervised anomaly detection : We assume here the availability of labeled data for a single class (generally for normal instances). This type of data allows to construct representative models for the normal behavior.
- Unsupervised anomaly detection : We do not assume that labeled data are available. This techniques are hence the most widely applicable.

Some classical methods

- **Unsupervised anomaly detection**
  - Without target variable to model : This is the framework where many methods have been proposed.
  - A first category of outlier detection methods is *density based*, where a distribution is used to fit the data (in a nonparametric way), and the outliers are defined from this distribution.
  - Another important category of outlier detection methods are *distance based* methods such as HCA, distance to the k nearest neighbors ..
  - In this course we focus on a *depth based* method (Isolation Forest) and *density based* methods (LOF and One Class SVM).

Outline

- Introduction
- Different aspects of an anomaly detection problem
- Focus on some anomaly detection algorithms
  - Isolation Forest
  - Local Outlier Factor
  - One Class SVM
- Conclusion

Isolation Forest

- Most model-based approaches to anomaly detection construct a profile of normal instances and determine data that do not correspond to the normal profile.
- The Isolation Forest algorithm is introduced in Liu et al. (ICDM 08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining).
- The principle is to isolate anomalies instead of profiling the normal behavior. The algorithm takes advantage from the fact that the anomalies are few and have very different values as normal instances from at least one variable.
- A tree structure is constructed to isolate every observation : anomalies are easier to isolate and thus will be isolated closer to the root of the tree while normal instances are isolated much further from the root.

Isolation Forest

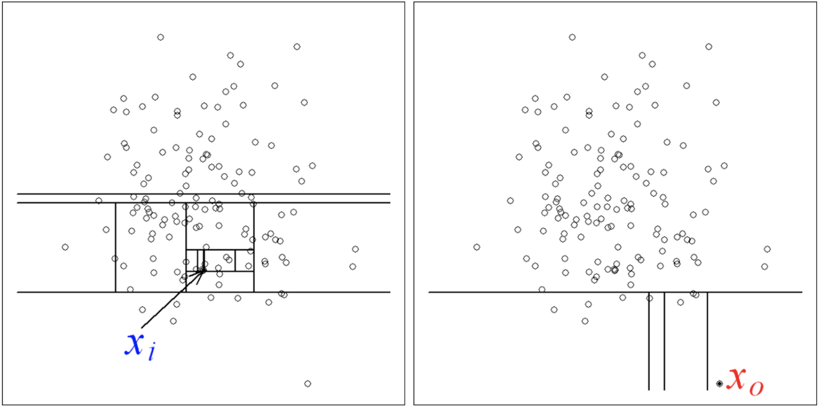


FIGURE – x_0 is easier to isolate than x_1. Source : towardsdatascience.com

Isolation Forest

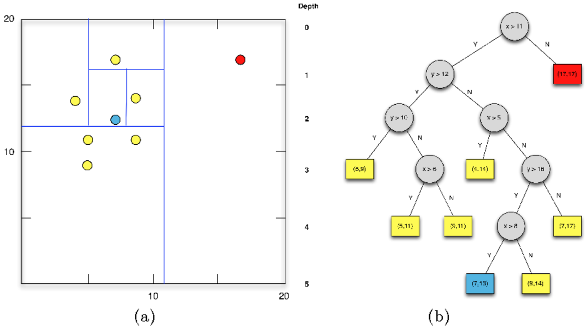


FIGURE – An iTree. Source : www.semanticscholar.org

Isolation Forest

- The tree is called Isolation tree or iTree.
- With the same principle as Random Forest, Isolation Forest or iForest builds an ensemble of iTrees. The instances that have short average path lengths in the iTrees are considered as anomalies.
- There are two parameters in the algorithm : the sub-sampling size used to built each tree and the number of trees.
- An advantage of this methods lies in the fact that it uses no distance or density measure to detect anomalies.
- This reduces the computation cost. iForest has a linear time complexity and a low memory requirements.

Isolation Forest

- The principle is to build a random tree, partitioning the observations until all the observations are isolated.
- The random partitioning produces shorter paths for anomalies since observations with highly distinguishable values for some variables are more likely to be isolated in early partitioning.
- An Isolation tree is defined as follows : given a data set $X = \{x_1, \ldots, x_n\}$ of n observations with d- dimensional distribution,
    - We recursively divide the data set X by selecting randomly at each node T of the tree a variable q and a split value p.
    - The test $q < p$ divides the data points of the node T into two son nodes T_l and T_r.
    - The subdivision is done until (i) the tree reached a height limit or (ii) all the points of the data set are isolated.

Isolation Forest

- The path length $h(x)$ of a point x is the number of edges the point x traverses from the root to its terminal node where the point is isolated.
- From this path length, we have to define an anomaly score. Since this quantity depends on the number n of observations, a proper normalization is necessary.
- The anomaly score $s(x, n)$ is defined by

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

where $E(h(x))$ is the average of $h(x)$ from a collection of isolation trees, and $c(n)$ is the harmonic number

$$c(n) = 2H(n-1) - 2(n-1)/n$$

with $H(i) = \ln(i) + 0.5772$ (Euler constant).

Isolation Forest

- $s(x,n)$ increases when $E(h(x))$ decreases.
  - When $E(h(x)) \to 0$, $s(x,n) \to 1$
  - When $E(h(x)) \to n-1$, $s(x,n) \to 0$
  - When $E(h(x)) \to c(n)$, $s(x,n) \to 1/2$
  - An observation x with $s(x,n)$ close to 1 is an anomaly.
  - An observation x with $s(x,n)$ smaller that $1/2$ can be considered as normal.
  - If all the observations have $s(x,n) \approx 1/2$, there is no anomaly in the data set.
- The value of $s(x,n)$ allows to rank the observations. Of course the cut-off for $s(x,n)$ between the observations declared as normal or anomalous is not obvious.
- A study of the distribution of $s(x,n)$ can help, as well as an idea on the proportion of anomalies in the data. One can also determine the top m anomalies for example.

Isolation Forest

Isolation Forest is a process in two steps :

- The first step (training) builds t isolation trees using t random subsamples of size ψ of the training set X. The complexity of this step is $O(t\psi\log(\psi))$.
- The second step (testing) passes the test instances through each iTree of an iForest to get an estimation of the expected path length $E(h(x))$ in order to compute an anomaly score for each observation of the test sample. The complexity of this step is $O(n_t t\log(\psi))$, where n_t is the size of the test sample.

Isolation Forest

- For high dimensional data, anomaly detection is even more difficult : distance-based methods are not appropriate since all the point are isolated in high dimension ; isolation Forest also suffers from the curse of dimensionality.
- A possible solution is to provide a pre-selection of the variables that are "candidate" for anomaly detection.
- For example, one can compute the Kurtosis (estimation of the fourth moment), which is sensitive to the presence anomalies and hence, is a good selector for the variables that may contain anomalies. It provides a ranking on the variables, and we can use only few selected variables for the Isolation Forest algorithm.

Outline

- Introduction
- Different aspects of an anomaly detection problem
- Focus on some anomaly detection algorithms
  - Isolation Forest
  - Local Outlier Factor
  - One Class SVM
- Conclusion

Local Outlier Factor

- The LOF is introduced in a paper by Breuning et al (2000) . The principle is to assign to each object a degree (the local outlier factor) of being an outlier.
- This is a density-based method. The LOF depends on how isolated the object is with respect to its neighborhood.
- The local outlier factor is based on the idea of "local density", where locality is given by the k nearest neighbors, whose distance is used to estimate the density.
- By comparing the local density of an object to the local densities of its neighbors, one can identify points that have a lower density than their neighbors, they are considered to be outliers.

Local Outlier Factor

LOF

We have n objects and for the sake of simplicity we assume that $\forall x_i, x_j, x_l,\ d(x_i, x_j) \neq d(x_i, x_l)$.
We consider $y, o' \in \{x_i, 1 \leq i \leq n\}$.

- We denote by $N_k(y)$ the set of the k nearest neighbors of y and by $d_k(y)$ the distance between y and its kth neighbour.
- We now define the *reachability-distance* of y and o' by

$$\text{r-dist}_k(y, o') = \max(d_k(o'), d(o', y)).$$

- Note that this is not a distance (not symmetric). This represents the distance between o' and y, lower bounded by $d_k(o')$ to get more stability. The k nearest neighbors of o' are at the same reachability-distance of o'.

Local Outlier Factor

LOF

- The *local reachability density* of y is defined by

$$lrd(y) = \frac{1}{\frac{1}{k}\sum_{o' \in N_k(y)} \text{r-dist}_k(y, o')}.$$

- This represents the inverse of the average reachability distance of y from its neigbors.
- The local reachability density of y is compared to the one of its k nearest neighbors to define the Local Outlier Factor
- The LOF is defined by

$$LOF_k(y) = \frac{1}{k} \sum_{o \in N_k(y)} \frac{lrd(o)}{lrd(y)}.$$

- k is an hyperparameter of the procedure, to be chosen by the data scientist.

Local Outlier Factor

- A value of the LOF approximately equal to 1 indicates that the object is comparable to its neighbors (and thus not an outlier), a value below 1 indicates a denser region, values significantly larger than 1 indicate outliers.

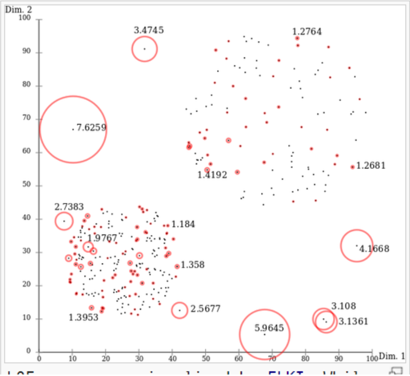
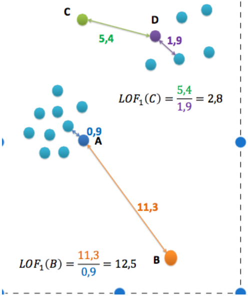


FIGURE – Local Outlier Factor. Source : https ://stats.stackexchange.com/

Local Outlier Factor

For the simple case $k = 1$, the following picture helps the understanding. In this case, $LOF(y)$ is simply the ratio between the distance of y to its nearest neighbor o' and the distance of o' to its nearest neighbor o.


C
D
5,4
1,9
$LOF_1(C) = \frac{5,4}{1,9} = 2,8$
0,9
A
11,3
$LOF_1(B) = \frac{11,3}{0,9} = 12,5$
B

Outline

- Introduction
- Different aspects of an anomaly detection problem
- Focus on some anomaly detection algorithms
  - Isolation Forest
  - Local Outlier Factor
  - One Class SVM
- Conclusion

One-class SVM

- This method has been introduced in the paper by Schölkopf et al. *Support Vector Method for Novelty Detection* (2000)
- The idea is to estimate the support of the probability density of the input observations.
- More precisely, the algorithm returns a function f which takes the value $+1$ in a "small" region capturing most of the data points and -1 elsewhere.
- Like for SVM's, the data are mapped into a feature space, associated to a kernel.
- The algorithm separates the data in the feature space from the origin with maximal margin.
- A point x is considered as an outlier if $f(x) = -1$.

One-class SVM

- We consider a training set $x_1, \dots, x_n \in \mathcal{X}$. Let Φ be a feature map $\Phi : \mathcal{X} \to F$ associated with the kernel $k(x,y) = \langle \Phi(x), \Phi(y) \rangle_F$, for example the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/c)$.
- The previous objectives are translated into the following optimization problem :

$$\begin{array}{cc} \displaystyle\min_{w \in F, \xi \in \mathbb{R}^n, \rho \in \mathbb{R}} & \displaystyle\frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_i \xi_i - \rho \\ \text{under the constraints} & \langle w, \Phi(x_i) \rangle_F \geq \rho - \xi_i,\ \xi_i \geq 0, \end{array}$$

  where $\nu \in]0,1[$ is an hyperparameter. The variables ξ_i represent the slack variables, they are penalized in the objective function.
- The decision function is

$$f(x) = sgn\left(\langle w, \Phi(x) \rangle_F - \rho\right).$$

One-class SVM

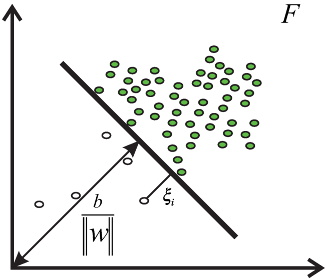


FIGURE – One Class SVM. Ref : Santiago-Paz et al., Entropy 2015, 17(9), 6239-6257

One-class SVM

- The parameter ν controls the trade-off between a small value of $\|w\|^2$(which corresponds to the maximization of the margin) and the slack variables.
- The dual problem is expressed as follows :

$$\min_{\alpha \in \mathbb{R}^n} \quad \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(x_i, x_j)$$

$$\text{under the constraints} \quad 0 \leq \alpha_i \leq \frac{1}{\nu n}, \sum_i \alpha_i = 1.$$

- The decision function is

$$f(x) = sgn\left(\sum_i \alpha_i k(x_i, x) - \rho\right).$$

One-class SVM

- One can prove that ν is an upper bound on the fraction of outliers.
- If the distribution of the data does not contain discrete components, under certain conditions on the kernel (satisfied by the Gaussian kernel), ν corresponds to the fraction of outliers with probability that tends to 1 when $n \to \infty$.

Outline

- Introduction
- Different aspects of an anomaly detection problem
- Focus on some anomaly detection algorithms
  - Anomaly detection with Random Forest
  - Isolation Forest
  - Local Outlier Factor
  - One Class SVM
- Conclusion

Conclusion

- Anomaly detection is an important problem with applications in a wide variety of domains.
- A key point is that it is not a well-formulated problem.
- Every formulation, depending on the application an the type of anomalies one aims to detect, may lead to appropriate detection algorithms.
- An important aspect on the subject is also the preprocessing on the data and the extraction of suitable features that will highlight anomalies (especially for complex data such as functional data or images).

References

Charu C. Aggarwal, *Outlier Analysis*, Springer (2013).

L. Breiman *Random Forests*, Machine Learning, **45** (2001), 5-32.

L. Breiman and A. Cutler. *Random Forests Manual v4.0*, Technical Report, UC Berkeley, 2003. Available online at : https ://www.stat.berkeley.edu/ breiman/

M. Breunig, H.P. Kriegel, T. Raymond and J. Sander, *LOF : Identifying Density-based Local Outliers*, Proceedings of the ACM Conference (2000), p. 93-104.

H. Caussinus and A. Ruiz-Gazen, *Classification and generalized PCA*, Springer 2007, https ://hal.archives-ouvertes.fr/hal-00635541.

V. Chandola, A. Banerjee and V. Kumar, *Outlier Detection : A survey*, ACM Computing Surveys 41(3) (2009).

F.T. Liu, K.M. Ting, and Z.H. Zhou, *Isolation Forest*, Proceedings of IEEE International Conference on Data Mining (2008), p. 413-422.

B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, J.C. Platt *Support vector method for novelty detection*, Advances in Neural Information Processing Systems 12, MIT PRESS, (2000), p. 582-588.