TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel and Prof. Dr. Kai Heinrich
Chair of Business Information Systems, esp. Intelligent Systems and Services

Data Science: Advanced Analytics

Anomaly Detection

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel and Prof. Dr. Kai Heinrich
Chair of Business Information Systems, esp. Intelligent Systems and Services

Anomaly Detection

Introduction

Introduction
Overview

Description: Detections of patterns that do not conform to „normal“ behavior

Application domains:

- Fraud Detection (e.g., credit card fraud, tax fraud, etc.)
- Intrusion Detection in cyber security
- Fault Detection in safety critical systems
- Military surveillance of enemies
- …

Importance of anomalies: Anomalies often are indicators for situations that require critical decisions

Example: An anomalous pattern in network traffic could indicate a hacker attack

Introduction
Overview

Anomalous patterns are often referred to as **anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities** or **contaminants** in different application domains

We will use the term **anomaly**

Main reference: Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. ACM Computing Surveys, 41(3), 1–58.

Distinction:

- Anomaly Detection deals with detecting interesting, non-normal patterns
- In contrast **noise removal** deals with detection and removal of unwanted **outliers** that are not of interest
- **Novelty detection** deals with new patterns that in contrast to anomalies can also be part of the „normal" behavior
- The methods used are fairly similar, however anomaly detection seems to offer the biggest range of methods

Introduction

Definition and Data Mining context

One possible definition:

> “An Outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980)
>
> Source: Hawkins D. 1980. Identification of Outliers. Chapman and Hall.

In the context of data mining:

- A portion of the observed data is generated by an unknown, anomalous process
- This process differs from the standard process

Until now:

- Assumption of one process that generates the data
- Data Mining models reflecting that process

Now: Interesting new observations that indicate differing mechanism behind data

Introduction

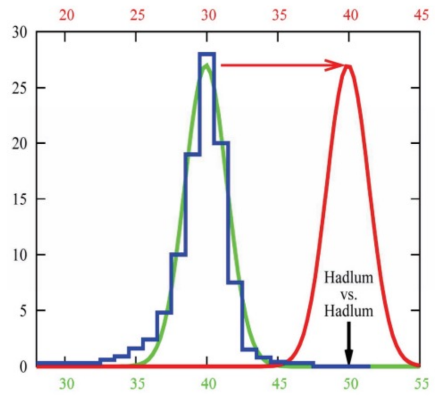
Example: Hadlum vs. Hadlum

The Case: Mr. Hadlum was filing for divorce because of adultery of his wife. The only evidence was the birth date of his child which was born 349 days after Mr. Hadlum shipped out for military service

The evidence: Mr. Hadlum argued that 349 days is far too anomalous compared to 280 average days of pregnancy, so he concluded it was not his child.

Question: Does the magnitude of difference between the standard process and the Hadlum case possibly indicate an anomaly ?

Hadlum v. Hadlum



- Green indicates the normal process (Mr. Hadlum = Father). We can see that the point in questions lies in a region very far right from this process
- Red indicates a possible anomalous process where Mr Hadlum would not be the father but someone else.

Introduction
Challenges

Defining a **normal region** which encompasses every possible normal behavior is very difficult.

When anomalies are the result of malicious actions, the malicious adversaries often **adapt themselves to** make the anomalous observations **appear like normal.**

In many domains normal behavior keeps **evolving** and a current notion of normal behavior might not be sufficiently representative in the future.

The exact notion of an anomaly is different for different application domains.

Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue.

Often the data contains **noise** which tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

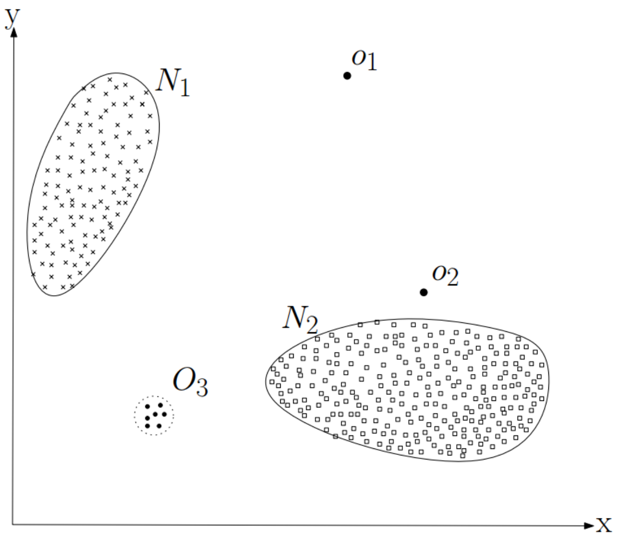
Anomaly Detection
Basic problem

Basic Models deal with anomalies as one-dimensional problems (see case Hadlum v. Hadlum)

But: Data is usually high dimensional

Goal: Infer the nature of the unknown process from the anomalous data

Basic idea of an Anomaly



- Regions N1 and N2 refer to data generated by „normal“ behavior
- Points o1 and o2 as well as region O3 could indicate anomalous behavior
- Problems:
  - What if we just do not have enough data of O3?
  - Is o2 really that far from „normal“?
  - Could o1 just be a noisy point?

Methods

Criteria for anomaly detection problem definition

There is **NO general** approach that works for all application domains

Different **problem definitions** depending on some **basic criteria**:

- Nature of data
- Labels
- Anomaly Type
- Output

Anomaly Detection methods can be found in many **research fields**:

- Machine Learning (Data Mining)
- Statistics
- Information Theory
- Spectral Theory
- …

Methods

Criteria: Nature of Input Data

Recall that input data are **attributes (variables)** measured for **data instances** (or **data objects**), e.g. measuring weight and blood pressure (attributes) of several patients (data objects) in a hospital

Scale: Attributes can be measured on different scale levels (**nominal, ordinal, interval, ratio**) ⇨ different techniques for different scale levels

Relationship between data instances: data objects can have different types of relationship to each other

- Sequential
- Spatial
- Graph data

⇨ Those relationships can be useful to determine anomalies

Methods

Criteria: Type of anomaly

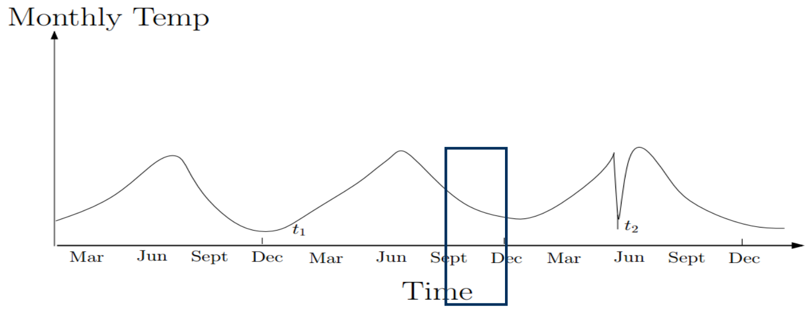
Point Anomalies:

- A single data object is considered to be not „normal“ in relation to the rest of the data
- Example: „Amount spend“ in credit fraud is very high compared to usual behavior

Contextual Anomalies:

- A single data point is only anomalous in a certain context
- The context in form of context variables sets the neighborhood and defines normal behavior
- E.g. observed temperature is normal during winter but are anomalous in June

Contextual Anomaly

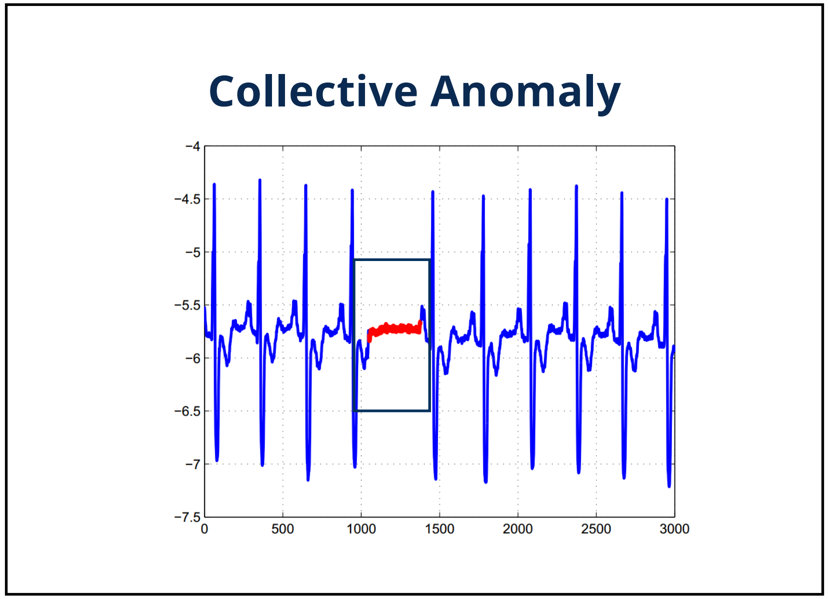


Methods

Criteria: Type of anomaly

Collective Anomalies:

- A collection of data objects is anomalous compared to the entire data set
- Elements of the collection might not be anomalous themselves but in combination they are
- **Example:** Low values in EKG for a long time; low values themselves are not abnormal but as a long lasting pattern they are



Methods

Criteria: Data labels and Output

To **classify** a data instance as normal or anomalous one can use **labels**:

- Labeling data is costly and expensive
- Getting labels for all kinds of anomalous behavior is nearly impossible
- Normal class easier to label

Depending on the availability of labeled data there are three approaches:

- **Supervised anomaly detection:** assumes labels for both classes
- **Semi-supervised:** only assumes labels for the normal class
- **Unsupervised:** does not assume labels

The **outputs** of an anomaly detection method can be:

- **Scores:** Scores how unusual the anomaly is compared to normal; ranking of the data objects possible
- **Labels:** Labels the data instance either anomalous or normal

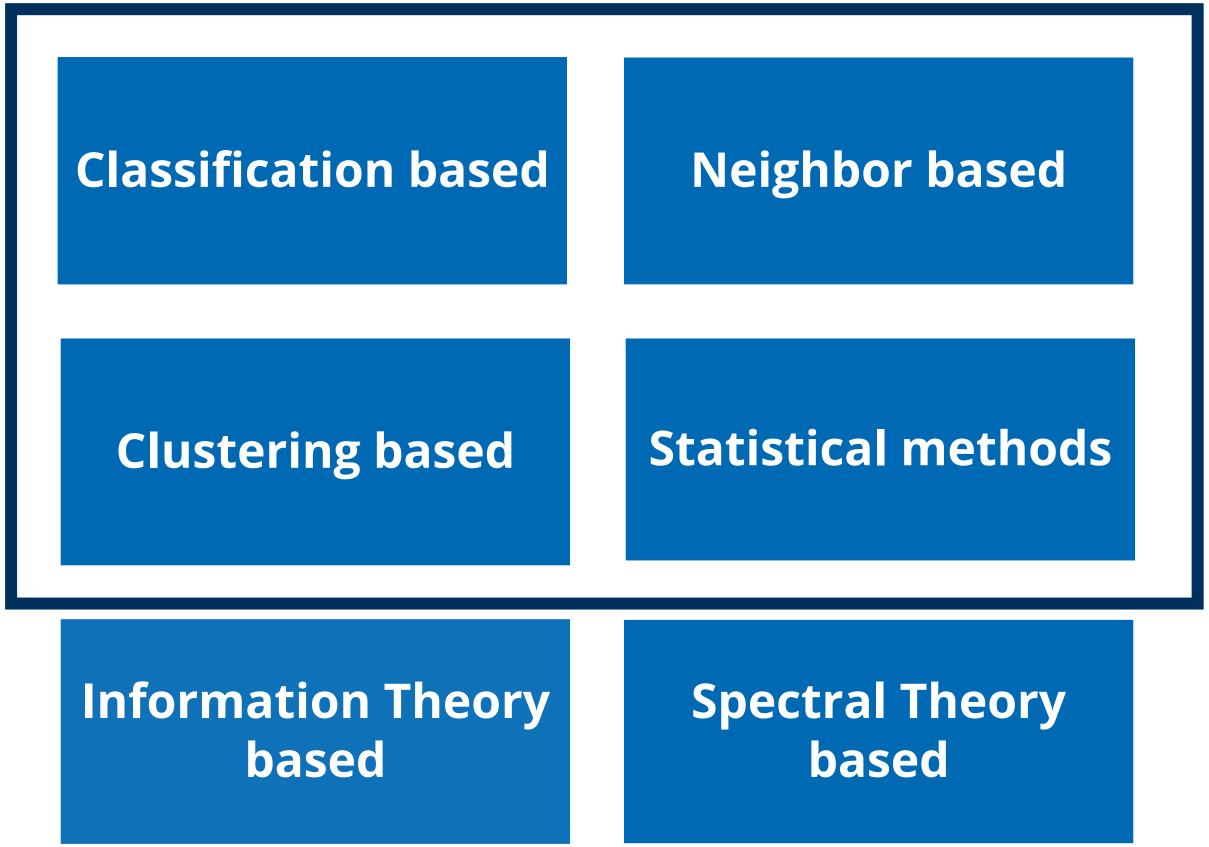
Prof. Dr. Alfred Benedikt Brendel and Prof. Dr. Kai Heinrich
Chair of Business Information Systems, esp. Intelligent Systems and Services

Anomaly Detection

Methods

Methods
Overview

Anomaly Detection Methods*



Classification based

Neighbor based

Clustering based

Statistical methods

Information Theory based

Spectral Theory based

*For Information Theory based and Spectral Theory based see [Chandola et al., 2009]

Classification based

Overview

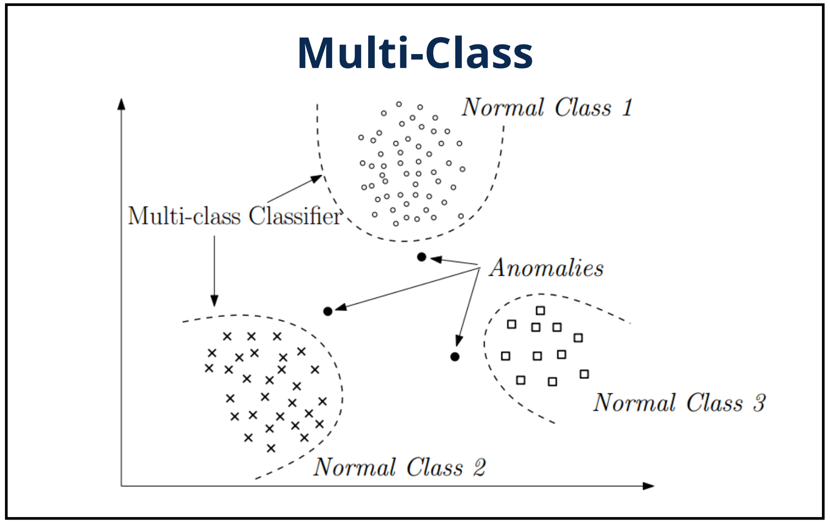
Supervised approach with „normal“ and „anomalous” classes

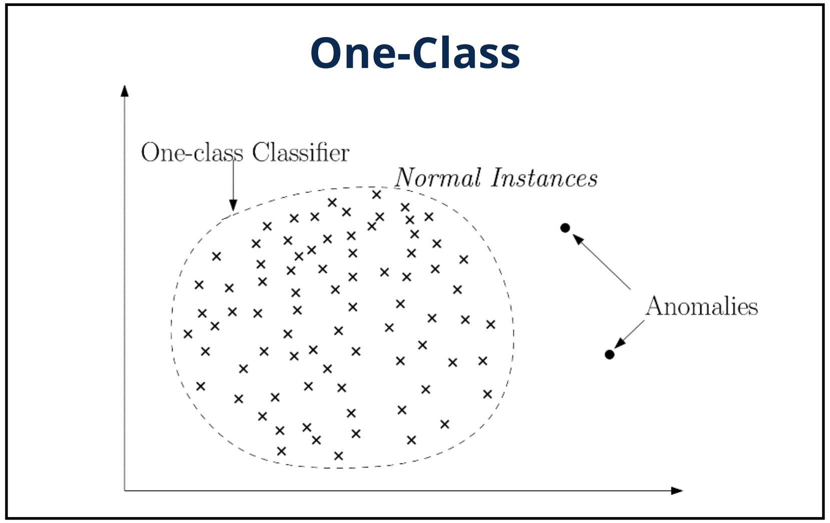
We assume we only know the „normal“ class

Two approaches:

- **Multi-Class:** Multi normal classes as labels ⇨ multiple boundaries
- **One-Class:** One normal class ⇨ one boundary

All points outside the boundaries are classified as „anomalous“





Classification based

One-class classification

From Wikipedia, the free encyclopedia

In machine learning, **one-class classification**, also known as **unary classification**, tries to *identify* objects of a specific class amongst all objects, by learning from a training set **containing only the objects of that class**. This is different from and more difficult than the traditional classification problem, which tries to *distinguish between* two or more classes with the training set containing objects from all the classes.

An example is the classification of the operational status of a nuclear plant as 'normal': In this scenario, there are (fortunately) few or no examples of catastrophic system states, only the statistics of normal operation are known.

The term **one-class classification** was coined by Moya & Hush (1996) and many applications can be found in scientific literature, for example outlier detection, **anomaly detection**, novelty detection.

A important component of one-class classification is that it uses only samples from the assigned class, so that a representative sampling is not strictly required for non-target classes.

Classification based

Multi-class classification

From Wikipedia, the free encyclopedia

In machine learning, **multiclass or multinomial classification** is the problem of classifying instances into one of the more than two classes (classifying instances into one of the two classes is called **binary classification**).

While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies.

Multiclass classification should not be confused with **multi-label classification**, where multiple labels are to be predicted for each instance.

Classification based

Algorithms

Neural Networks:

- **Multi-Class:** Network is trained on normal data, test data is either accepted or rejected
- **One-Class:** Using Replicator Neural Networks, the input data is replicated by a number of hidden layers (usually three). The reconstruction error resembles the anomaly score

Decision Tree:

- Builds rules based on normal behavior; if test object does not follow any rules it is declared anomalous
- Anomaly Score is inverse value of the leaf confidence of the best rule for each test point (= the fraction of correctly classified values)

Support Vector Machine:

- Learns complex regions; If a point is outside the region it is declared an anomaly
- **Special Case:** One Class SVM

Classification based

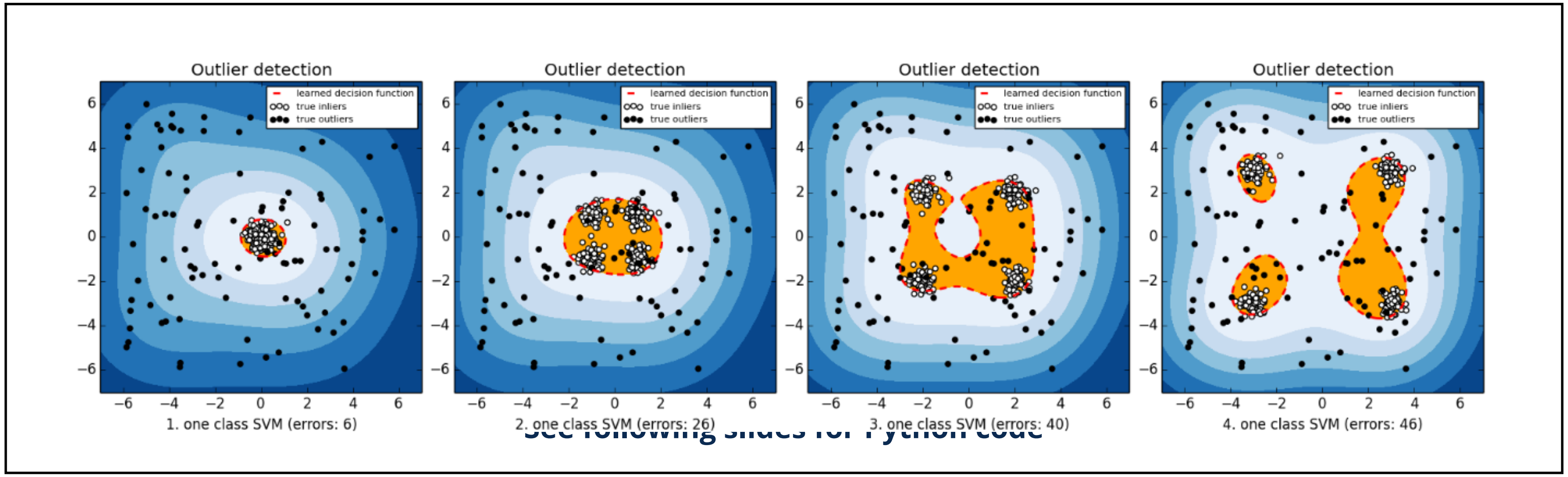
One-class SVM

Idea: Generate an area of „normal“ example around the data limited by the SVM margin

Labels: +1 for a small area around the normal examples and -1 elsewhere

Algorithm: Using common SVM approach with only one example in the -1 class which is the origin (called one-class SVM or ν-SVM)

Configuration: new parameter ν enables trade off between false classification (if ν approaches 0 there is a high penalty for false classes) and number of points inside the margin (if ν approaches 1)



1. one class SVM (errors: 6)

2. one class SVM (errors: 26)

3. one class SVM (errors: 40)

4. one class SVM (errors: 46)

See following slides for Python code

Classification based

Advantages and disadvantages

Advantages:

Classification based techniques, especially the multi-class techniques, can make use of powerful algorithms that can distinguish between instances of anomalous and normal cases

The testing phase of classification based techniques is fast

Disadvantages:

Multi-class classification based techniques rely on availability of accurate labels for various normal classes, which is often not possible.

Classification based techniques assign a **label** to each test instance, which can also become a disadvantage when a meaningful anomaly **score** is desired for the test instances

Neighbor based

Overview

Idea: Normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors.

We need to measure the distance between data object to construct the neighborhood.

We can use distance measures introduced in the cluster lectures.

Remember: Distance measures should be chosen according to the scale levels of variables.

Two main approaches in anomaly detection:

- Approaches that use **k-nearest neighbor** distances
- Approaches that calculate **relative density scores** for each object with regard to the neighborhood

Neighbor based
k-nearest neighbor

Idea: The anomaly score of a data object is given by the distance to its k-nearest neighbor

- Parameter *k* has to be set
- A threshold on the score is usually needed (when is the anomaly score high enough so that the data point is considered to be abnormal?)

Other Definition: Calculate the number *k* of neighbors that are within distance *d* of a data object

The inverse term *1/d* is used as anomaly score

Other ideas:

- Hypergraph based Outlier test (HOT): Using connectedness of a graph for categorical and / or mixed data scales
- Hilbert Space Filling Curves: Improvements over algorithmic complexity issues

Neighbor based

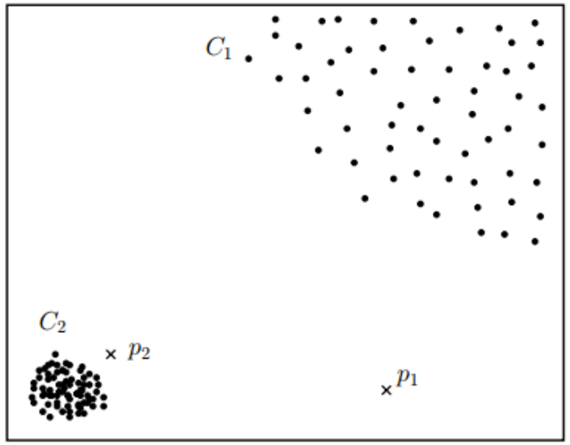
Relative density: Problem

Idea: If a data instance lies in a neighborhood with a low density it is more likely to be anomalous

Distance to k-nearest neighbor is now used as the radius of a hypersphere

Problem: This basic idea performs poorly if there are regions with varying densities:

Density based neighborhoods



- Normal regions C1 and C2 have different densities
- C2 is much more dense than C1
- Problems:
  - Distance between instances in C1 and nearest neighbors is bigger than the distance between the nearest neighbors of C2 and p2 for every instance in C2
  - As a result p2 will not be considered anomalous
  - p1 might be considered anomalous

Neighbor based
Relative density: LOF

Local Outlier Factor (LOF): Takes average density of data object AND neighbors into account

- Find the K-nearest neighbors $N_k(A)$ of each point A in dataset.
- Calculate K-distance (*k–dist(A)*) as max distance of k nearest neighbor(s)
- Next we calculate the reachability-distance:

$$rd_k(A,B) = \max\{k\!-\!dist(B), d(A,B)\}$$

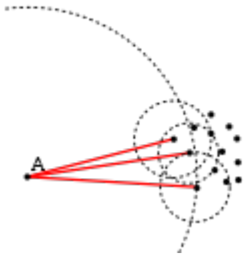
- Calculate the local reachability density

$$lrd_k\ (A) = 1/\left(\frac{\sum_{B\in N_k(A)} rd_k(A,B)}{|N_k(A)|}\right)$$

- Calculate LOF-Score: If the **LOF-Score** is significantly greater **than 1** then it indicates an anomaly.

$$LOF_k\ (A) = \frac{\sum_{B\in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|}$$

Idea of LOF



to compare the local density with others

Neighbor based
LOF: Example

Example Data:

- 2D Data (=2 Variables) with 4 rows= 4 points
- a(0, 0), b(0, 1), c(1, 1), d(3, 0)

Assumptions:

- k=2
- Distance used= Manhattan Distance

Step 1: Calculate all the distances between the points

- dist(a, b) = 1
- dist(a, c) = 2
- dist(a, d) = 3
- dist(b, c) = 1
- dist(b, d) = 3+1=4
- dist(c, d) = 2+1=3

Neighbor based
LOF: Example

Step 2: Calculate $dist_k(o)$ (=distance between a point o and its k -th nearest neighbor)

$dist_2(a) = dist(a, c) = 2$ (c is the 2nd nearest neighbor)

$dist_2(b) = dist(b, a) = 1$ (a/c is the 2nd nearest neighbor)

$dist_2(c) = dist(c, a) = 2$ (a is the 2nd nearest neighbor)

$dist_2(d) = dist(d, a) = 3$ (a/c is the 2nd nearest neighbor)

Step 3: Collect all nearest neighbor points $N_k(o)$ to a point o, by checking which $dist(o', o) \leq dist_k(o)$, for all other points o'

$$N_2(a) = \{b, c\}$$
$$N_2(b) = \{a, c\}$$
$$N_2(c) = \{b, a\}$$
$$N_2(d) = \{a, c\}$$

— dist(a, b) = 1

— dist(a, c) = 2

— dist(a, d) = 3

— dist(b, c) = 1

— dist(b, d) = 3+1=4

— dist(c, d) = 2+1=3

Neighbor based

LOF: Example

Step 4: Calculate local reachability distance $lrd_k(o)$ for each point o

Remember:

- $lrd_k\ (o) = 1/\left(\frac{\sum_{o\prime \in N_k(o)} rd_k(o,o\prime)}{|N_k(o)|}\right)$
- We need $|N_k(o)|$ (cardinality of the set $N_k(o)$ → how many points are in the neighbourhood of o
- We need the reachability distances $rd_k(o, o')$ of o to every point o' in the neighbourhood of o

Example for point a:

- $|N_k(a)| = 2$
- $rd_2(a, b) = \max\{dist_2(b), dist(a, b)\} = \max\{1,1\} = 1$
- $rd_2(a, c) = \max\{dist_2(c), dist(a, c)\} = \max\{2,2\} = 2$
- $lrd_2(a) = \frac{|N_k(a)|}{rd_2(a,b)+rd_2(a,c)} = \frac{2}{1+2} = 0.667$

Neighbor based

LOF: Example

Step 4: Calculate local reachability distance $lrd_k(o)$ for each point o

Similar for the other points:

- $lrd_2(b) = \frac{|N_k(b)|}{rd_2(b,a)+rd_2(c,a)} = \frac{2}{2+2} = 0.50$
- $lrd_2(c) = \frac{|N_k(c)|}{rd_2(b,c)+rd_2(a,c)} = \frac{2}{1+2} = 0.667$
- $lrd_2(d) = \frac{|N_k(d)|}{rd_2(a,d)+rd_2(c,d)} = \frac{2}{3+3} = 0.333$

Step 5: Calculate LOF

Remember:

- $LOF_k\ (A) = \frac{\sum_{B \in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{|N_k(A)|} = \sum_{o\prime \in N_k(o)} lrd_k(o') \sum_{o\prime \in N_k(o)} rd_k(o', o)$

Neighbor based

LOF: Example

Step 5: Calculate LOF scores for all points:

- $LOF_2(a) = (lrd_2(b) + lrd_2(c))(rd_2(b,a) + rd_2(c,a) = (0.5 + 0.667)(1 + 2) = 3.501$
- $LOF_2(b) = (lrd_2(a) + lrd_2(c))(rd_2(a,b) + rd_2(c,b) = (0.667 + 0.667)(2 + 2) = 5.336$
- $LOF_2(c) = (lrd_2(b) + lrd_2(a))(rd_2(b,c) + rd_2(a,c) = (0.5 + 0.667)(1 + 2) = 3.501$
- $LOF_2(d) = (lrd_2(a) + lrd_2(c))(rd_2(a,d) + rd_2(c,d) = (0.667 + 0.667)(3 + 3) = 8.004$

Conclusion: Point d is top outlier with highest LOF score

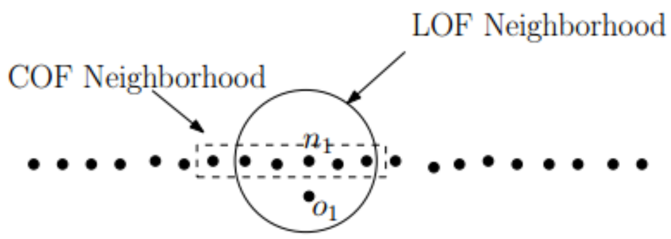
Neighbor based

Relative density: LOF Variations

Variations:

- **ODIN:** simpler version of LOF
- **MDEF:** variation of LOF – based on the standard deviation of the local densities
- **COF:** variation with an incremental calculation of the k-neighborhood
- Other variations to improve computation time and support for other data types (e.g. sequence data)

LOF vs. COF neighborhood



Neighbor based
Advantages

Unsupervised nature ⇨ does not make any assumptions regarding the generative distribution for the data.

Semi-supervised techniques **perform better** than unsupervised techniques in terms of missed anomalies.

Adapting nearest neighbor based techniques to a different data type is straightforward.

Neighbor based
Disadvantages

For unsupervised techniques, if the data has normal instances that do not have enough close neighbors or if the data has anomalies that have enough close neighbors, the **technique fails** to label them correctly, resulting in missed anomalies.

For semi-supervised techniques, if the normal instances in test data do not have enough similar normal instances in the training data, the **false positive rate** for such techniques is high.

The **computational complexity** of the testing phase is also a significant challenge since it involves computing the distance of each test instance with all instances belonging to either the test data itself, or to the training data, to compute the nearest neighbors.

Performance of a nearest neighbor based technique greatly relies on a distance measure, defined between a pair of data instances, that can effectively distinguish between normal and anomalous instances.

Clustering based
Overview

Idea 1: Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster:

Requires cluster method that does not put every data point into clusters

E.g.: DBSCAN, ROCK, WaveCluster, FindOut

Idea 2: Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid:

Common used methods are SOM, K-means, Expectation Maximization

Semi-supervised approach can be used to improve results

Idea 1 and 2 are **not working well** when anomalies form clusters themselves

Idea 3: Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters

Commonly used: **FindCBLOF**

CBLOF analogous to LOF captures cluster size as well as centroid distance based on clustering results

Clustering based
Advantages

Clustering based techniques can operate in an **unsupervised mode**.

Such techniques can often **be adapted to other complex** data types by simply plugging in a clustering algorithm that can handle the particular data type.

The **testing phase** for clustering based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant.

Clustering based
Disadvantages

Performance of clustering based techniques is highly dependent on the effectiveness of clustering algorithm in capturing the cluster structure of normal instances.

Many techniques detect anomalies as a **by-product of clustering**, and hence are not optimized for anomaly detection.

Several clustering algorithms force every instance to be assigned to some cluster.

Several clustering based techniques are effective only when the anomalies do not form significant clusters among themselves.

The **computational complexity** for clustering the data is often a bottleneck.

Statistical methods
Overview

Idea: Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Parametric approaches:

- The normal data is assumed to be generated from a probability distribution
- The inverse of the probability of an observation to be generated from such a distribution is considered as anomaly score ⇨ small probability means high chance of anomaly
- Statistical tests are widely used

Non-parametric methods:

- No probability distribution is assumed a-priori
- Instead the distribution form is estimated based on the data

Statistical methods

Parametric methods: Gaussian model

Gaussian assumption:

Assumes a normal distribution to fit the data

Parameters of the distribution are estimated based on data using statistical inference (e.g. MLE)

The distance of a data object to the estimated mean is the anomaly score

Different tests:

Simple $3-\sigma$ rule: Every point that lies beyond 3 standard deviations is considered an anomaly

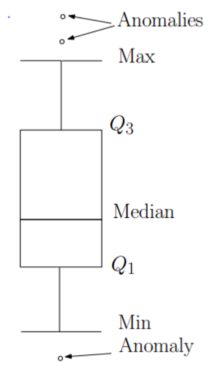
Box-Plot-Rule:

- Every point k* IQR above Q3 or
  k*IQR below Q1 is considered an outlier
- The boxplot therefore gives left bound (Min)
  and right bound (Max) thresholds for anomalies

Univariate tests: Grubbs test, t-test

Multivariate tests: Hotelling t²-test, Slippage detection, Rosner test, Dixon test

Outlier Boxplot



Statistical methods

Parametric methods: other methods

Regression based:

(1) Regression is estimated

(2) for each data object the residual score resembles the anomaly score

Some criteria used to detect anomalies in a regressions setting:

- AIC: anomalies influence regression results
- Robust regression: should be used to avoid anomaly influences ⇨ residuals for anomalies are considerably larger in this setting
- ARIMA/ARMA: detection of anomalies in a time series setting

Mixture distribution based: Two approaches - Different distributions for normal data and anomalies or mixed distribution for only the normal class

Statistical methods
Non-parametric methods

Histogram based:

- (1) Build a histogram to resemble the distribution f of the data
- (2) Anomaly score based on the height of the bin in which the data object falls; simpler version: if it does not fit to any bin it is considered an anomaly
- Definition for bin size is the key factor:
  - If there are small bins only, many normal cases fall in rare bins
  - If there are large bins only, many rare cases end up in normal bins
- Multivariate version: Combine univariate scores to a data set score
- Variants are often times used in Intrusion detection (ALAD and PHAD)

Kernel Function based:

- Similar to parametric anomaly detection but the probability function is estimated based on the data
- E.g.: Parzen Window Estimation

Statistical methods

Advantages

If the **assumptions** regarding the underlying data distribution hold true, statistical techniques provide a statistically justifiable solution for anomaly detection.

The anomaly score provided by a statistical technique is associated with a **confidence interval**, which can be used as additional information while making a decision regarding any test instance.

If the **distribution estimation** step is **robust** to anomalies in data, statistical techniques can operate in a unsupervised setting without any need for labeled training data.

Statistical methods

Disadvantages

The key disadvantage of statistical techniques is that they rely on the **assumption** that the data is generated from a particular distribution. This assumption often does not hold true, especially for high dimensional real data sets.

Even when the statistical assumption can be reasonably justified, there are several hypothesis test statistics that can be applied to detect anomalies; choosing the right one often poses a problem.

Histogram based techniques are relatively simple to implement, but a key shortcoming of such techniques for multivariate data is that they are **not able** to capture the **interactions between different attributes**.

Anomaly Detection
Summary

Even though there exist several types of multivariate anomalies, most of the approaches are using **point anomalies** as an assumption.

Each of the large number of anomaly detection techniques have their unique strengths and weaknesses ⇨ **situation based application**.

Anomaly detection techniques typically assume that anomalies in **data are rare** when compared to normal instances ⇨ not always the case (e.g. anomalous virus traffic is more frequent).

Every anomaly detection method has an idea in form of assumptions behind it ⇨ this can be used to check the fitness of that approach with regard to a specific topic.

Most of the **methods are commonly known** methods like SVM, Clustering, Kernel density estimation, etc. that are used in another context.

Prof. Dr. Alfred Benedikt Brendel and Prof. Dr. Kai Heinrich

Chair of Business Information Systems, esp. Intelligent Systems and Services

Thank you for your attention