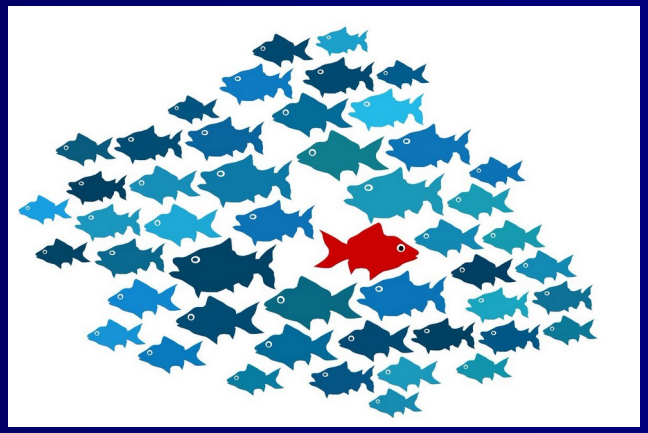
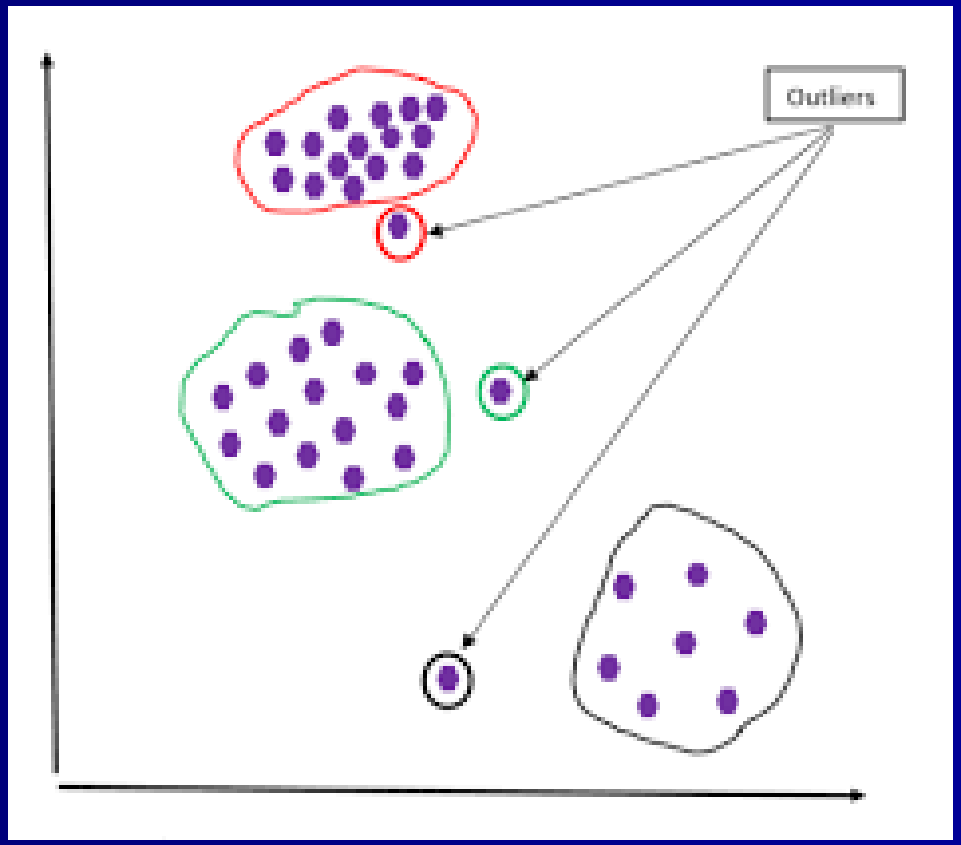
ONE-CLASS CLASSIFICATION
– OUTLIER DETECTION –





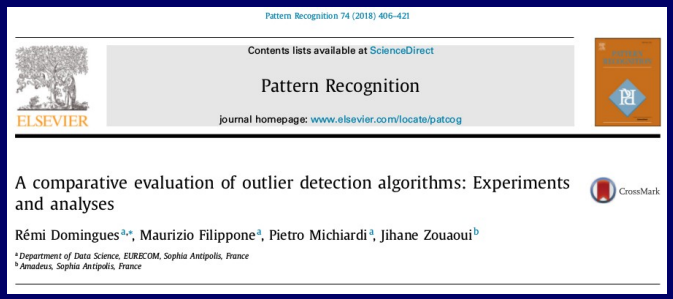
OUTLINE

- The terms
- One-Class Classification
- Outlier detection

One-class classification: Concept learning in the absence of counter-examples.
DMJ Tax - 2002 - elibrary.ru
Degree: Dr. DegreeYear: 2001 Institute: Technische Universiteit Delft (The Netherlands) Publisher: Print partners Ipskamp, Capitool 25, Postbus 333, 7500 AH Enschede, The Netherlands. This thesis treats the problem of one-class classification. It starts with an …
☆ 99 Citado por 1338 Artículos relacionados Las 5 versiones

- Other scenarios: anomaly detection, novelty detection

- Main “one-class classification” algorithms


Pattern Recognition 74 (2018) 406–421
Contents lists available at ScienceDirect
Pattern Recognition
journal homepage: www.elsevier.com/locate/patcog
A comparative evaluation of outlier detection algorithms: Experiments and analyses
Rémi Domingues[a,*], Maurizio Filippone[a], Pietro Michiardi[a], Jihane Zouaoui[b]
[a] Department of Data Science, EURECOM, Sophia Antipolis, France
[b] Amadeus, Sophia Antipolis, France

- References and software

Efficient **Outlier Detection** in **Text** Corpus Using Rare Frequency and Ranking

WA Mohotti, R Nayak - ACM Transactions on Knowledge Discovery from …, 2020 - dl.acm.org

… known to face difficulties to deal with higher dimensions, inherent to **text** data due … Density-based clustering methods such as DBSCAN can naturally **detect outliers** in the dataset … Several clusters-based **outlier detection** methods have been proposed focusing the tightness of the …

[HTML] Integrating aspect analysis and **local outlier factor** for intelligent review spam detection

L You, Q Peng, Z Xiong, D He, M Qiu… - Future Generation …, 2020 - Elsevier

… Aspect rating LOF (AR-LOF). In this subsection, we transform review spam detection into an **outlier** detection problem and employ the **local outlier factor** (LOF) algorithm to address it … First, previous **outlier** detection methods captured only certain kinds of **outliers**, since they …

[HTML] Unusual customer response identification and visualization based on **text mining** and **anomaly detection**

S Seo, D Seo, M Jang, J Jeong, P Kang - Expert Systems with Applications, 2020 - Elsevier

Abstract The Vehicle Dependability Study (VDS) is a survey study on customer satisfaction for vehicles that have been sold for three years. VDS data analytics plays an important role in the vehicle development process because it can contribute to enhancing the brand image …

A **Text Mining**-Based **Anomaly Detection** Model in Network Security

M Kakavand, N Mustapha, A Mustapha… - Global Journal of …, 2015 - computerresearch.org

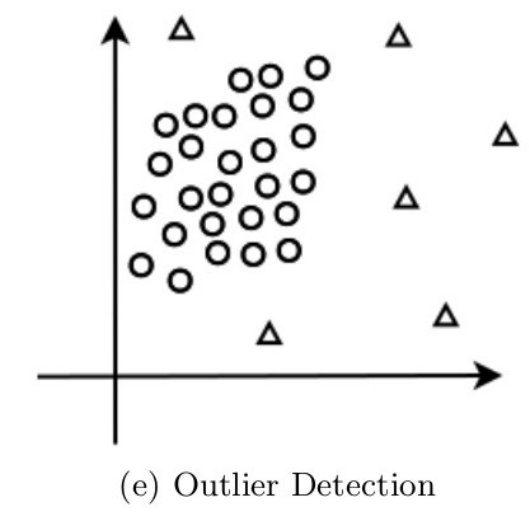
Anomaly detection systems are extensively used security tools to **detect** cyber-threats and attack activities in computer systems and networks. In this paper, we present **Text Mining**-Based **Anomaly Detection** (TMAD) model. We discuss n-gram **text** categorization and focus …

Cleaning Out Web **Spam** by Entropy-Based Cascade **Outlier Detection**

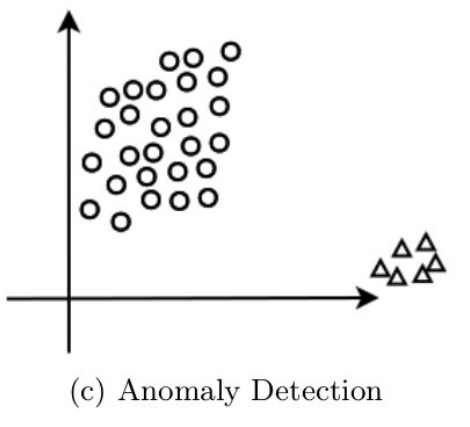
S Wei, Y Zhu - International Conference on Database and Expert …, 2017 - Springer

… achieved good experimental results based on the co-training model [3]. As the **spamming** has become … **spam** pages are of poor quality and have short life cycle, because the **spammers** want to … of an outgoing link, anchor text of links) could be helpful for discovering more **spam** …

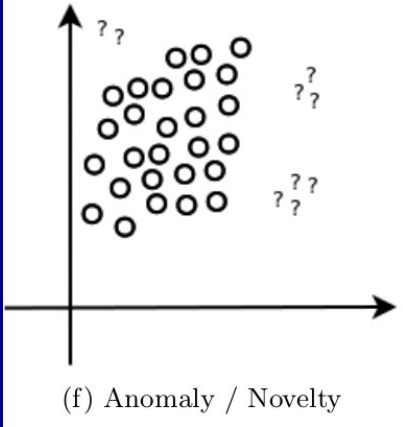
OUTLIER – ANOMALY – NOVELTY



(e) Outlier Detection



(c) Anomaly Detection

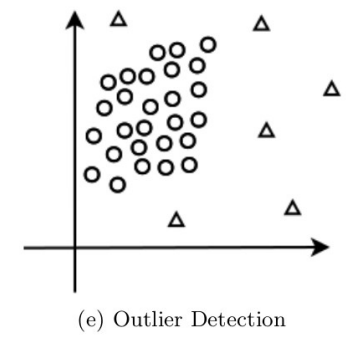


(f) Anomaly / Novelty

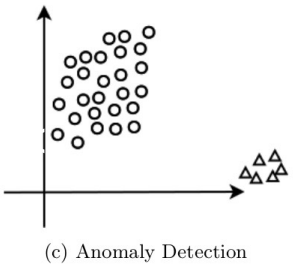
- "Normal" category → common, majority

- Outlier samples→ sparse, minoritary, ~~category~~
- Anomaly samples → minoritaria, categoría ~ imbalanced supervised learning
- Novelty samples → anomaly, unknown in training time

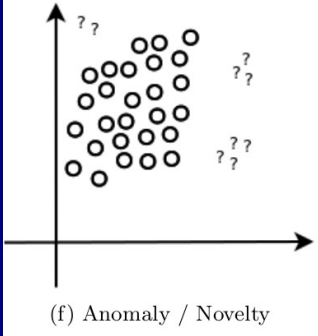
OUTLIER – ANOMALY – NOVELTY



(e) Outlier Detection



(c) Anomaly Detection



(f) Anomaly / Novelty

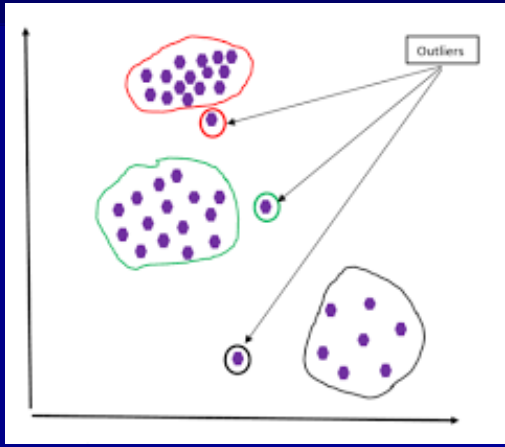
Artificial Intelligence Review (2020) 53:3575–3594
https://doi.org/10.1007/s10462-019-09771-y

Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework

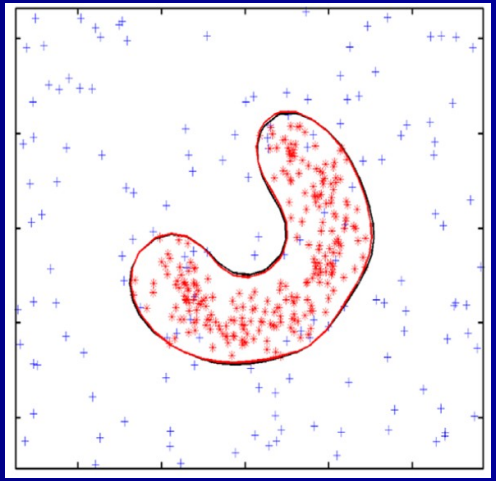
Ander Carreño[1] · Iñaki Inza[1] · Jose A. Lozano[1,2]

OUTLIER ~~ ONE-CLASS CLASSIFICATION

- Outlier samples → sparse, minoritary, ~~category~~
- Outlier scenario → multi-class scenario + more than one class

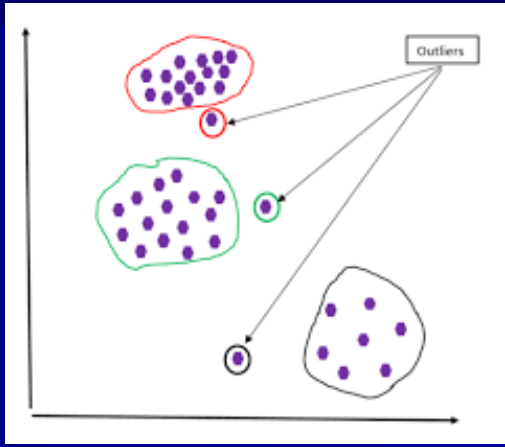


- One-class classification → single class
- Single class → model its boundary + isolate the "rest"
- Non-single-class samples → non-modeled

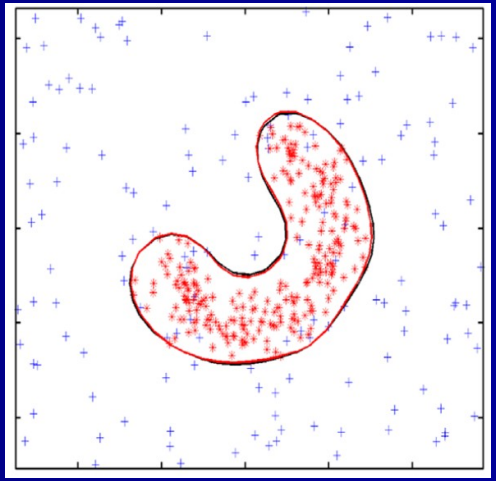


OUTLIER ~~ ONE-CLASS CLASSIFICATION

- Outlier samples → sparse, minoritary, ~~category~~
- Outlier scenario → multi-class scenario + more than one class



- One-class classification → single class
- Single class → model its boundary + isolate the "rest"
- Non-single-class samples → non-modeled



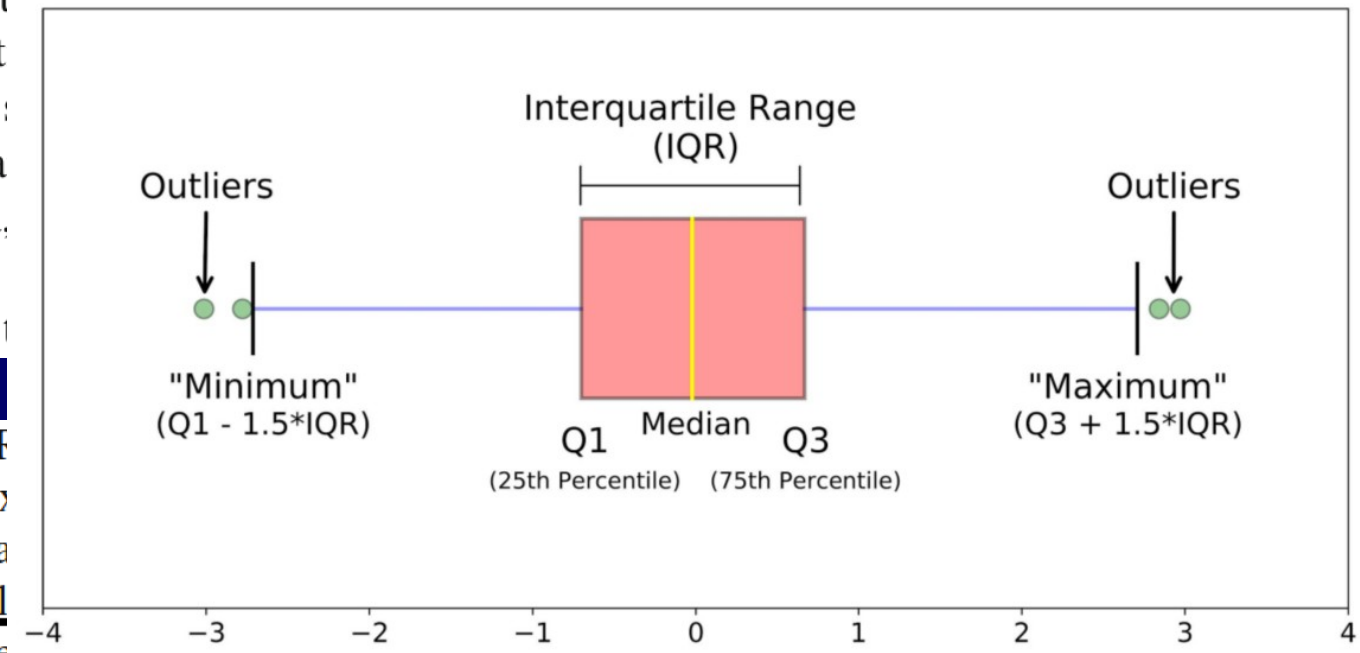
OUTLIER DETECTION
– THE UNIVARIATE APPROACH –

Whenever the goal is to identify univariate outliers, the statistical methods are among the simplest ones. Assuming a Gaussian distribution and learning the parameters from the data, parametric methods identify the points with low probability as outliers. One of the methods used to spot such outliers is the boxplot method, introduced by Tukey (1977). Based on the first quartile ($Q1$), the third quartile ($Q3$) and the interquartile range ($IQR = Q3 - Q1$) of the data, it determines that the interval $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$ contains 99.3% of data. Therefore, points outside that interval are considered as mild outliers, and points outside the interval $[Q1 - 3 * IQR, Q3 + 3 * IQR]$ are considered extreme outliers.

Regarding the unsupervised learning approach, we have used a boxplot-based method (boxplotEns) as follows: we generate five boxplots, one for each attribute, on the training data. For each new observation, we check if the value of each attribute is considered to be an outlier by the respective boxplot. The observation is considered to be an Abnormal cycle if at least one of the five attributes values is signalled as an outlier. This approach corresponds to an ensemble of boxplots to deal with multivariate data.

OUTLIER DETECTION
– THE UNIVARIATE APPROACH –

Whenever the goal is to identify univariate outliers, the statistical methods are among the simplest ones. Assuming a Gaussian distribution and learning the parameters from the data, paramet methods used to d on the first qua $- Q1$) of the data, s 99.3% of data. d points outside ers.



method (box training data to be an outl cycle if at least one of the five attributes' values is signaled as an outlier. This approach corresponds to an ensemble of boxplots to deal with multivariate data.

OUTLIER DETECTION DATASETS – BENCHMARKS –

- *What do you mean by "outlier"?*

- ***Multi-dimensional point datasets*** [* ← ours]
- Time series graph datasets for event detection
- Time series point datasets (uni/multi -variate time series)
- Cyber-attack scenarios – security datasets
- Crowded scene video for anomaly detection

ISOLATION FOREST

- Compute "isolation score" ***per sample***

- Construct a tree per sample
- Random splits on attribute values
- → isolates the sample from the rest
- → "outliers easy to isolate..."

- Path length from root to node
- ~ "isolation score" = "outlierness"

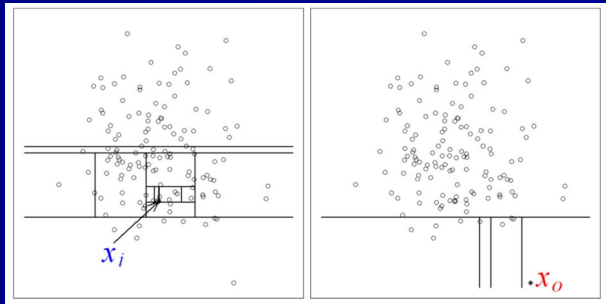
- "low path length" ~ "high outlierness"
- → easy to isolate point

- → graph "outlierness" values → threshold

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

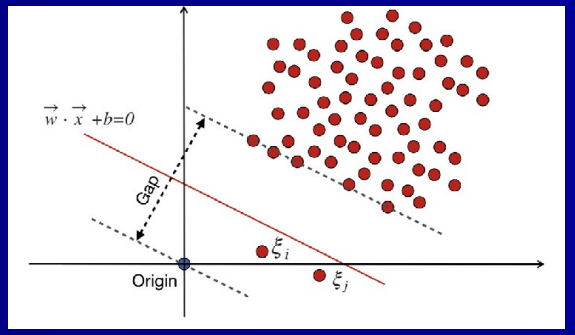


(a) Isolating x_i (b) Isolating x_o

```
# package with benchmark datasets
library(mlbench)
# Census data for 506 Boston houses
data("BostonHousing", package = "mlbench")
# Package with IsolationForest implementation
library(solitude)
# Empty tree structure
iso <- isolationForest$new()
# Learn the IsolationTree for our data
iso$fit(BostonHousing)
p <- iso$predict(BostonHousing)
print(p)
sort(p$anomaly_score)
plot(density(p$anomaly_score))
# based on the plot, decide the cut-off point:
# indexes of samples with Outlierness > 0.63
which(p$anomaly_score > 0.63)
```

OneClass SVMs – OCVSM

- Learn a SVM with single-class samples
- Map to higher dimension space
- Separating hyperplane
- Maximize margin between origin and data
- Outliers → points outside boundary



Support Vector Method for Novelty Detection

Bernhard Schölkopf*, Robert Williamson§,
Alex Smola§, John Shawe-Taylor†, John Platt*

* Microsoft Research Ltd., 1 Guildhall Street, Cambridge, UK
§ Department of Engineering, Australian National University, Canberra 0200
† Royal Holloway, University of London, Egham, UK
* Microsoft, 1 Microsoft Way, Redmond, WA, USA
bsc/jplatt@microsoft.com, Bob.Williamson/Alex.Smola@anu.edu.au, john@dcs.rhbnc.ac.uk

```r
library(e1071)
# Daily air quality measurements in New York,
# May to September 1973.
# https://stat.ethz.ch/R-manual/R-devel/library/
# datasets/html/airquality.html
data(airquality)
df <- airquality

# all variables to be numerical

#train a SVM one-classification model
model <- svm(df, y=NULL,type='one-classification')


print(model)
summary(model) #print summary

# test on the whole set
# TRUE values mean suspect outliers
pred <- predict(model, df)
which(pred==TRUE)
```

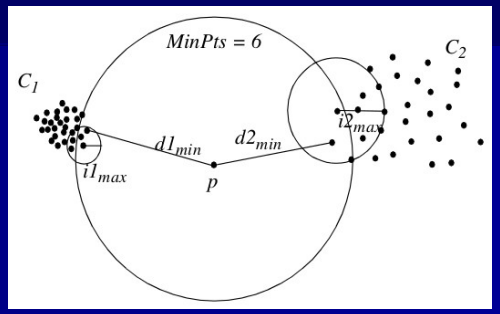
LOCAL OUTLIER FACTOR - LOF

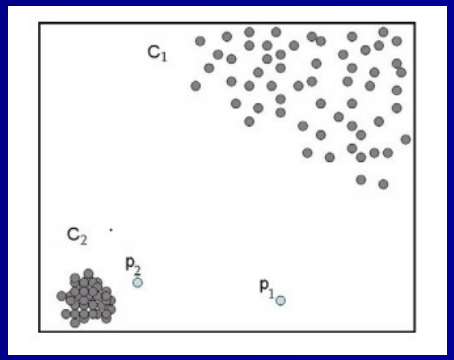
- Distance-based algorithm

- To decide “outlier”
- → by local neighborhood
- → by local density

- Parameter → k, number of neighbours
- Calculate the neighborhood

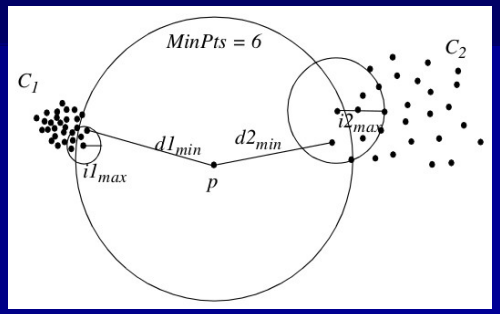
- Outlier → defined “locally”
- Outlierness → compute density of its local k-neighborhood






```r
library(DDoutlier)
# 1860 Daily Closing Prices of Major European Stock Indices
# https://stat.ethz.ch/R-manual/R-devel/library/
# datasets/html/EuStockMarkets.html
data("EuStockMarkets")
colnames(EuStockMarkets)
# calculate "outlierness" score, by LOF
outlierness = LOF(dataset=EuStockMarkets, k=5)
# assign an index to outlierness values
names(outlierness) <- 1:nrow(EuStockMarkets)
sort(outlierness, decreasing=TRUE)
hist(outlierness)
which(outlierness > 2.0)
```

LOCAL OUTLIER FACTOR - LOF



- Distance-based algorithm

- To decide “outlier”
- → by local neighborhood
- → by local density

- Parameter → k, number of neighbours
- Calculate the neighborhood

- Outlier → defined “locally”
- Outlierness → compute density of its local k-neighborhood

Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000

LOF: Identifying Density-Based Local Outliers

Markus M. Breunig†, Hans-Peter Kriegel†, Raymond T. Ng‡, Jörg Sander†

† Institute for Computer Science
University of Munich
Oettingenstr. 67, D-80538 Munich, Germany
{ breunig | kriegel | sander }
@dbs.informatik.uni-muenchen.de

‡ Department of Computer Science
University of British Columbia
Vancouver, BC V6T 1Z4 Canada
rng@cs.ubc.ca

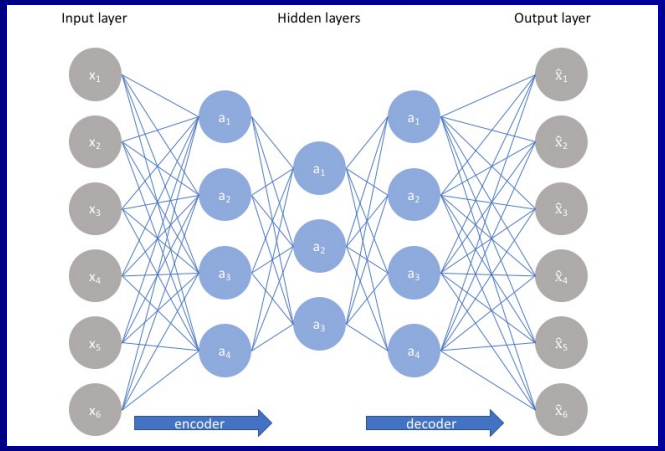
```
library(DDoutlier)
# 1860 Daily Closing Prices of Major European Stock Indices
# https://stat.ethz.ch/R-manual/R-devel/library/
# datasets/html/EuStockMarkets.html
data("EuStockMarkets")
colnames(EuStockMarkets)
# calculate "outlierness" score, by LOF
outlierness = LOF(dataset=EuStockMarkets, k=5)
# assign an index to outlierness values
names(outlierness) <- 1:nrow(EuStockMarkets)
sort(outlierness, decreasing=TRUE)
hist(outlierness)
which(outlierness > 2.0)
```

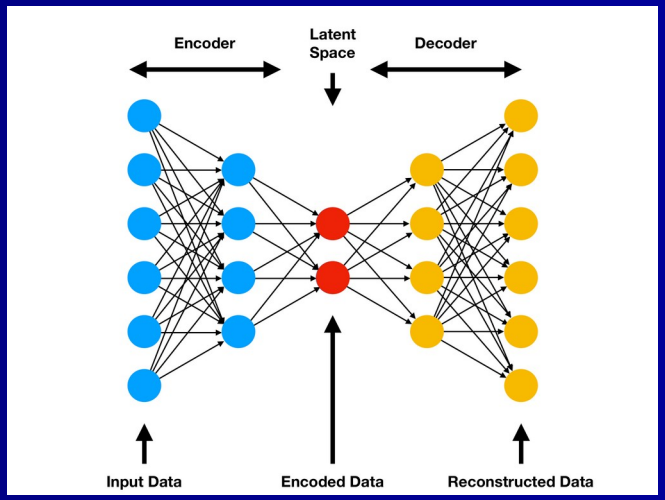
AUTOENCODERS
– DEEP LEARNING –

- Learn “encoded” data representation
- Reducing to non-linear dimensions in hidden layers
- {Encode + Decode} 1-class data

- Check for anomalies
- Does the autoencoder “reconstruct” the input data in the output?
- → “reconstruction error”
- → high value indicative of outlierness

- Hidden layers' features
- Non-linear, compact representation
- → learn with them a supervised model?





AUTOENCODERS – DEEP LEARNING –

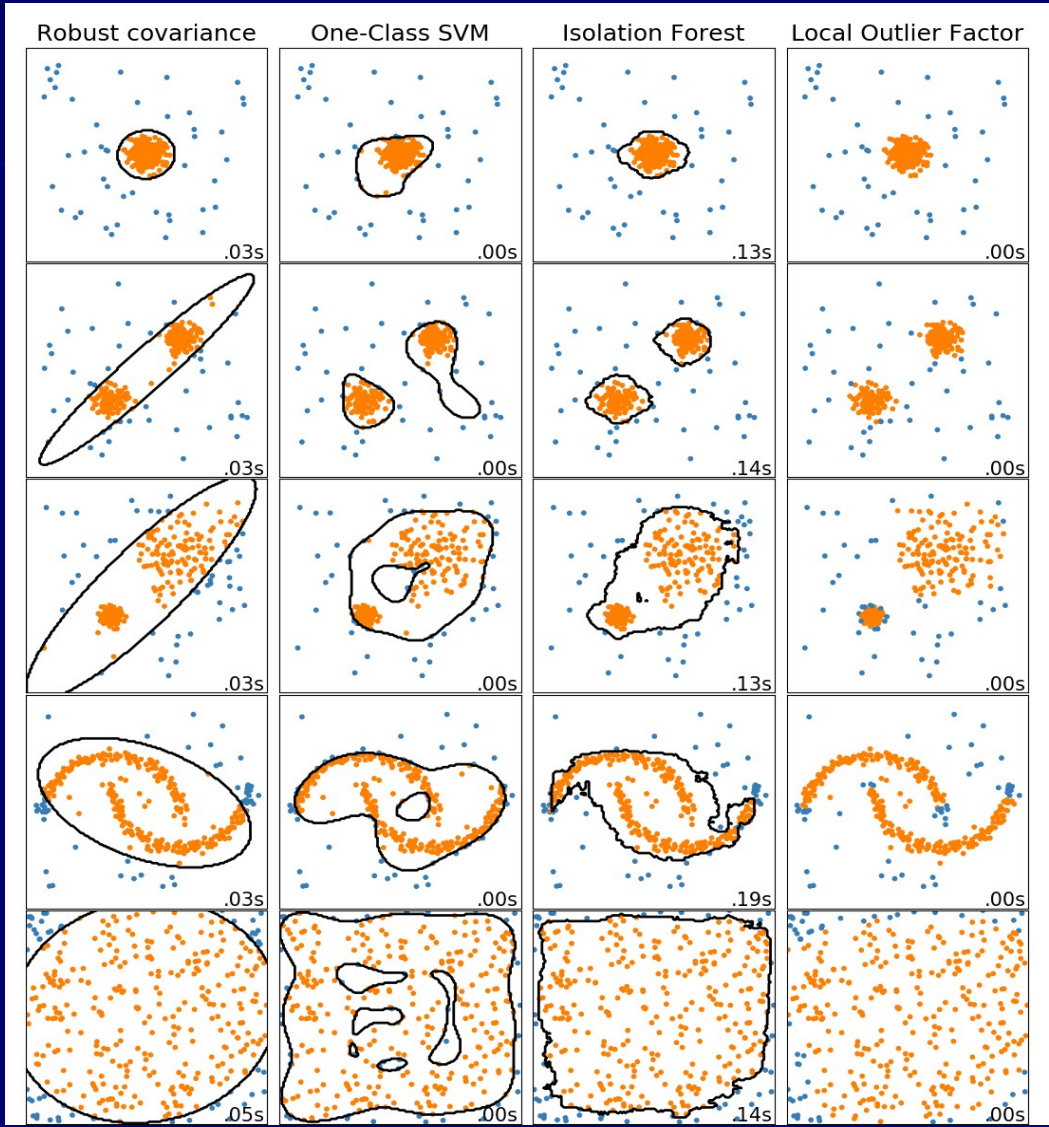
- Learn representation of data
- Reducing to non-linear dimensions in hidden layers
- {Encode + Decode} 1-class data

- Check for anomalies
- Does the autoencoder “reconstruct” the input data in the output?
- → “reconstruction error”
- → high value indicative of outlierness

- Hidden layers' features
- Compact, non-linear representation
- → learn with them a supervised model?

```
library(h2o)
h2o.init()
prostate_path = system.file("extdata",
                "prostate.csv", package = "h2o")
prostate = h2o.importFile(path = prostate_path)
colnames(prostate)
dim(prostate)
# learn autoencoder with 2 hidden layers of  10 units each
autoencoder_model = h2o.deeplearning(x = 3:9,
                    training_frame = prostate, autoencoder =TRUE,
                    hidden = c(10, 10), epochs = 5)
# features in the autoencoder's first hidden layer
deep_features_layer1 = h2o.deepfeatures(autoencoder_model,
                       prostate, layer=1)
# further supervised models can be trained with these features
head(deep_features_layer1)
# reconstruction error per sample ~ outlierness indicative
reconstruction_error = h2o.anomaly(autoencoder_model, prostate)
head(reconstruction_error)
reconstruction_error = as.data.frame(reconstruction_error)
plot(sort(reconstruction_error$Reconstruction.MSE),
     main='Reconstruction Error')
which(reconstruction_error > 0.15)
```

ONE-CLASS CLASSIFICATION



Pattern Recognition 74 (2018) 406–421

Contents lists available at ScienceDirect

Pattern Recognition

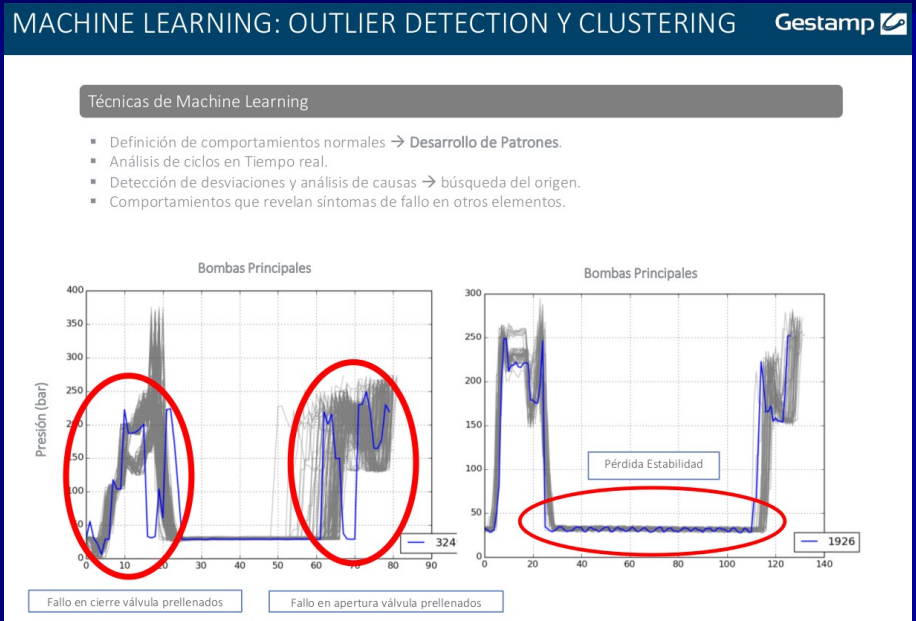
journal homepage: www.elsevier.com/locate/patcog

A comparative evaluation of outlier detection algorithms: Experiments and analyses

Rémi Domingues[a,*], Maurizio Filippone[a], Pietro Michiardi[a], Jihane Zouaoui[b]

[a] Department of Data Science, EURECOM, Sophia Antipolis, France
[b] Amadeus, Sophia Antipolis, France

“BASQUE” APPLICATION INDUSTRY 4.0



- “Machine-tool” manufacturers
- Non-availability of “failure-class data”

- Predictive maintenance – “early prediction”
- “Do not arrive to failure and avoid machine stop”

Efficient **Outlier Detection** in **Text** Corpus Using Rare Frequency and Ranking

WA Mohotti, R Nayak - ACM Transactions on Knowledge Discovery from …, 2020 - dl.acm.org

… known to face difficulties to deal with higher dimensions, inherent to **text** data due … Density-based clustering methods such as DBSCAN can naturally **detect outliers** in the dataset … Several clusters-based **outlier detection** methods have been proposed focusing the tightness of the …

[HTML] Integrating aspect analysis and **local outlier factor** for intelligent review spam detection

L You, Q Peng, Z Xiong, D He, M Qiu… - Future Generation …, 2020 - Elsevier

… Aspect rating LOF (AR-LOF). In this subsection, we transform review spam detection into an **outlier** detection problem and employ the **local outlier factor** (LOF) algorithm to address it … First, previous **outlier** detection methods captured only certain kinds of **outliers**, since they …

[HTML] Unusual customer response identification and visualization based on **text mining** and **anomaly detection**

S Seo, D Seo, M Jang, J Jeong, P Kang - Expert Systems with Applications, 2020 - Elsevier

Abstract The Vehicle Dependability Study (VDS) is a survey study on customer satisfaction for vehicles that have been sold for three years. VDS data analytics plays an important role in the vehicle development process because it can contribute to enhancing the brand image …

A **Text Mining**-Based **Anomaly Detection** Model in Network Security

M Kakavand, N Mustapha, A Mustapha… - Global Journal of …, 2015 - computerresearch.org

Anomaly detection systems are extensively used security tools to **detect** cyber-threats and attack activities in computer systems and networks. In this paper, we present **Text Mining**-Based **Anomaly Detection** (TMAD) model. We discuss n-gram **text** categorization and focus …

Cleaning Out Web **Spam** by Entropy-Based Cascade **Outlier Detection**

S Wei, Y Zhu - International Conference on Database and Expert …, 2017 - Springer

… achieved good experimental results based on the co-training model [3]. As the **spamming** has become … **spam** pages are of poor quality and have short life cycle, because the **spammers** want to … of an outgoing link, anchor text of links) could be helpful for discovering more **spam** …

EXERCISE

- Choose a publication → describing a previous method
- Read its abstract

- Find a software package which develops it
- Parameters of the method?

- Choose a supervised dataset (e.g. spambase)
- Choose one of its classes (e.g. “non-spam e-mails”)
- Apply the one-class method over it
- Be careful !! → methods may only work with numerical features
- → remove the class !!

- Graph “outlierness” distribution → cut-off point to decide outliers
- Are there suspicious outliers within this class e-mails?