Detection of Outliers for the Inspection of Potential Fraud Cases

Rita P. Ribeiro (rpribeiro@dcc.fc.up.pt)

Workshop on Machine Learning and Big Data

BANCO DE PORTUGAL
EUROSISTEMA

[dcc]
DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

LIAAD
INESCTEC
LABORATÓRIO ASSOCIADO

U.PORTO

Outline

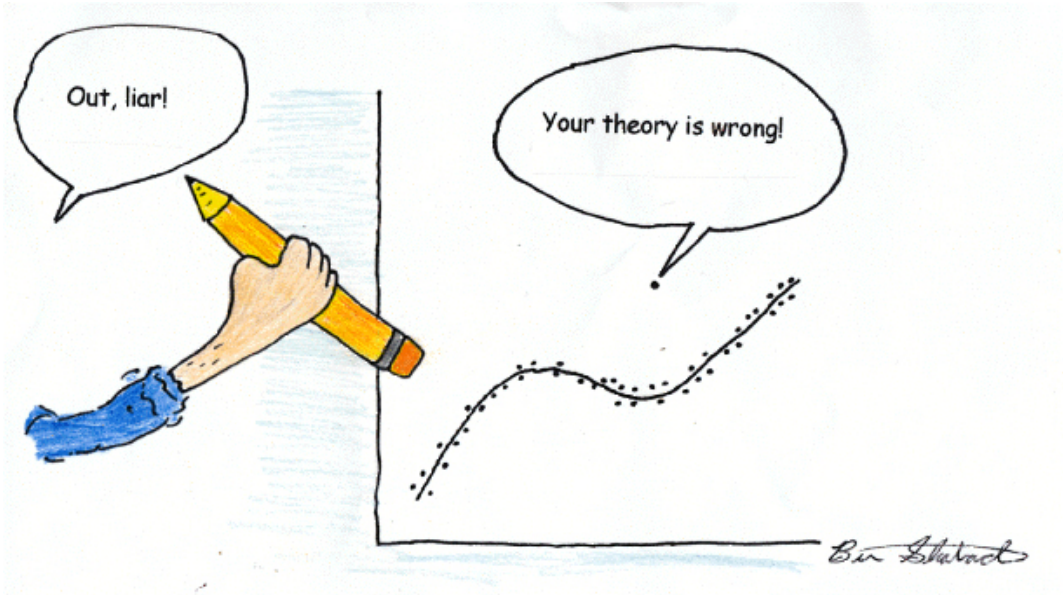
1. Outlier Detection Problem
2. Case Study: Detection of Potential Fraud Cases
3. Summary

Motivation

- Most of data mining tasks focus on creating a model of the "normal" patterns in the data, extracting knowledge from what is common (e.g. frequent patterns).
- Still, rare patterns can also give us some import insights about data.
- Depending on the goal, those insights can be even more interesting/critical than the "normal" patterns.

What is an Outlier?

- *"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)



What is an Outlier? (cont.)

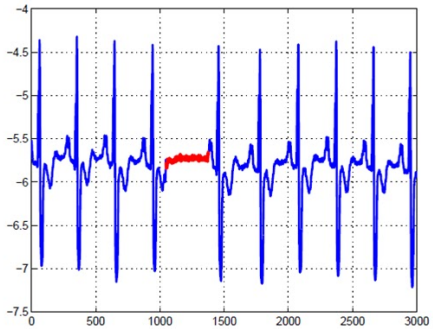
- Outliers represent patterns in data that do not conform to a well defined notion of normal.

- Initially, outliers were considered errors and their identification had data cleaning purposes.

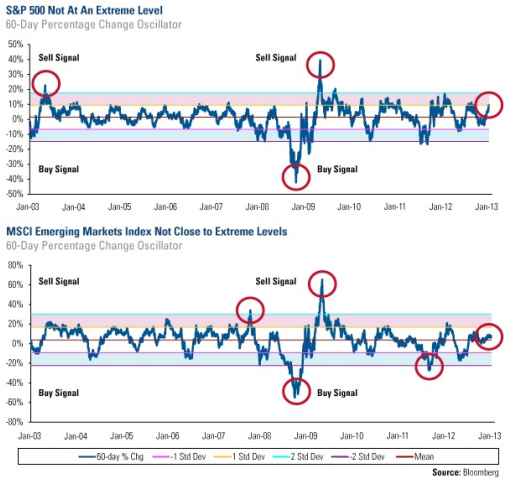
- However, they can represent truthful deviation of data.
- For some applications, they represent critical information, which can trigger preventive or corrective actions.

Where can Outliers occur?

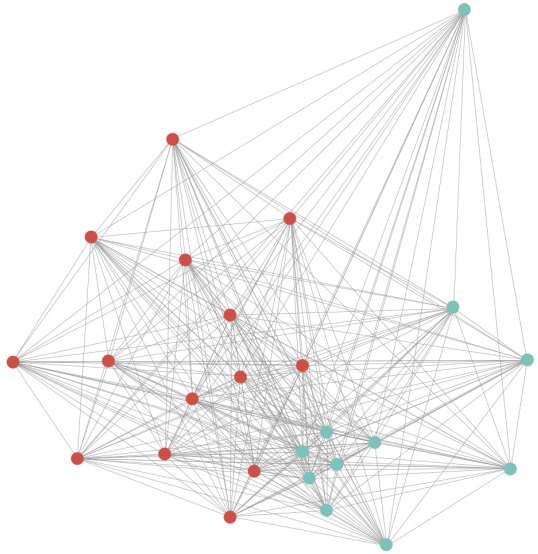
Medical Analysis



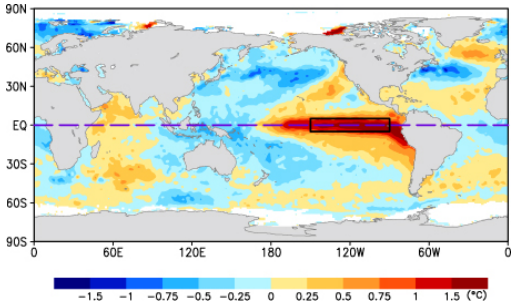
Financial Markets



Social Network Analysis

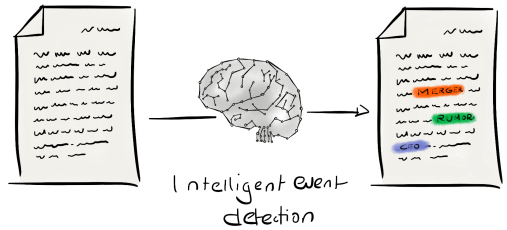


Anomalous Weather Patterns



Fraud Detection

Event Detection in Text/Social Media



Applications of Outlier Detection

- Medical Sensor / Imaging Diagnostics
- Anomalous Events in Environmental Sciences
- Quality Control and Fault Detection Applications
- Event Detection in Text and Anomalous Activity in Social Media
- Network Intrusion and Security Applications
- Financial Applications
  - Credit Card Fraud
  - Insurance Claim Fraud
  - Extreme Stock Market Returns
  - Unusual Interaction between Financial Entities
  - Accounting Fraud on Financial Reports

Challenges of Outlier Detection

- Define every possible "normal" behaviour is hard.
- The boundary between normal and a outlying behaviour is often not precise.
- There is no general outlier definition; it depends on the application domain.
- It is difficult to distinguish real meaningful outliers from simple random noise in data.
- The outlier behaviour may evolve with time.
- Malicious actions adapt themselves to appear as normal.
- Inherent lack of known labeled outliers for training/validation of models.

Key Aspects of Outlier Detection Problem

- Nature of Input Data
- Type of Outliers
- Intended Output
- Learning Task
- Performance Metrics

Nature of Input Data

- Each data instance has:
  - One attribute (univariate)
  - Multiple attributes (multivariate)
- Relationship among data instances:
  - None
  - Sequential / Temporal
  - Spatial
  - Spatio-temporal
  - Graph
- Dimensionality of data

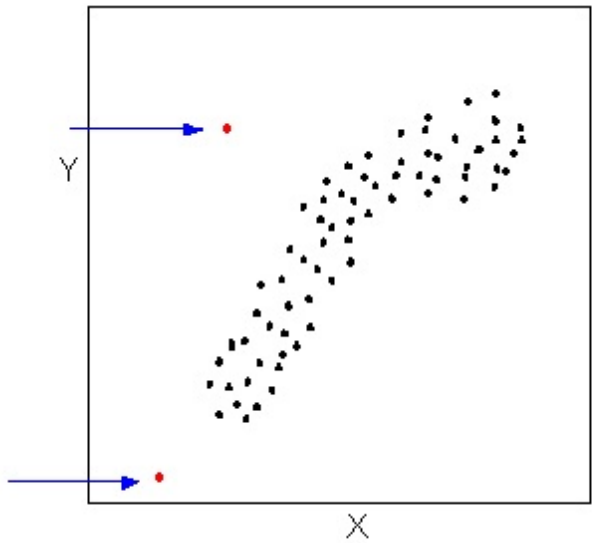
Types of Outliers

- Point (or Global) Outlier
- Contextual Outlier
- Collective Outlier

Types of Outliers (cont.)

Point Outlier

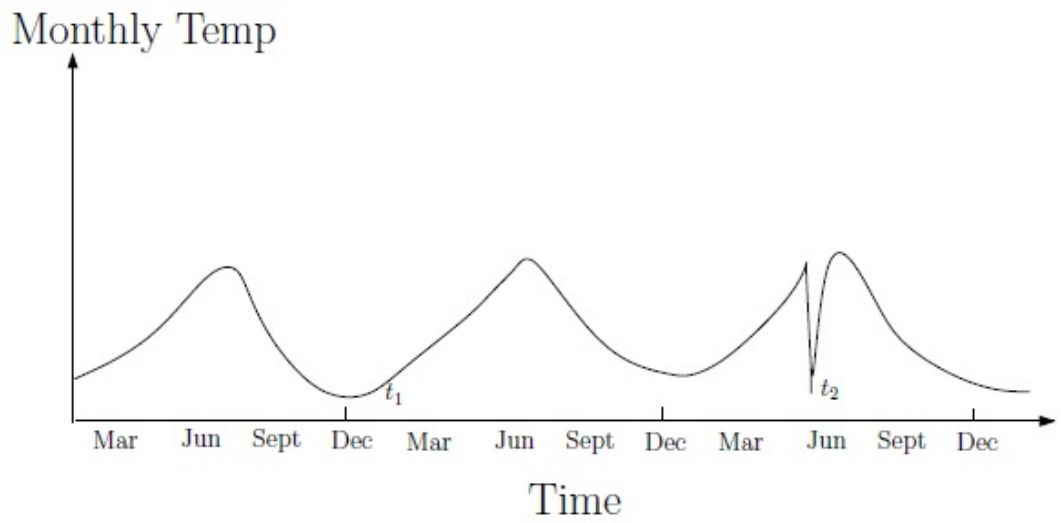
An instance that, individually or in small groups, is very different from the rest of the instances.



Types of Outliers (cont.)

Contextual Outlier

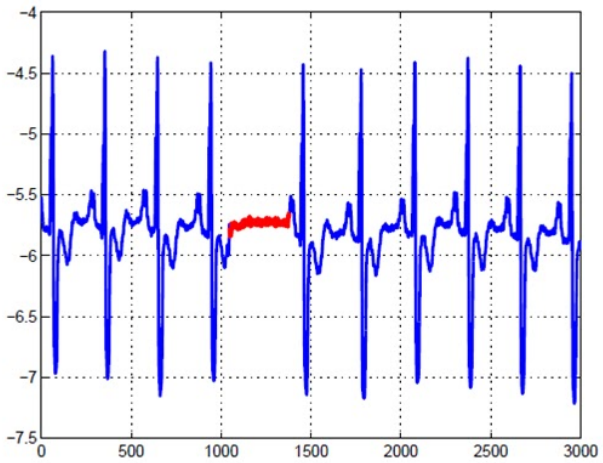
An instance that, when considered within a context, is very different from the rest of the instances.



Types of Outliers (cont.)

Collective Outlier

An instance that, even though individually may not be an outlier, inspected in conjunction with related instances and with respect to the entire data set is an outlier.

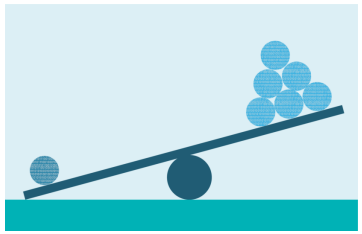


Intended Output

- Assign a label/value: identification normal or outlier instance
- Assign a score: probability of being an outlier
  - allows the output to be ranked
  - requires the specification of a threshold

Learning Task: Supervised Outlier Detection

- Data set has instances with both normal and outlier classes/values on the target variable;
- Imbalance learning problem:
  - training set has an imbalanced distribution;
  - the preference is not uniform over all target variable values;
  - the learning algorithm should focus on the rare cases;

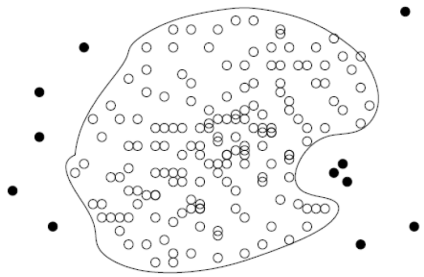


[source: https://datascience.aero]

- Limitations:
  - effectiveness depends on the reliability of target values;
  - in real-life applications, such data is hard to obtain;
  - cannot detect unknown or emerging outliers.

Learning Task: Semi-Supervised Outlier Detection

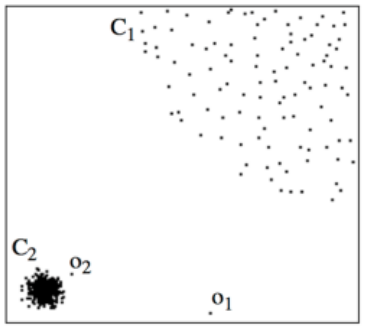
- Data set has instances of normal behaviour;
- Build a prediction model to the normal behaviour and identify any deviations from this behaviour as outliers.



- Limitations:
  - requires labeled instances for normal behaviour;
  - previously unseen normal data may be identified as an outlier;

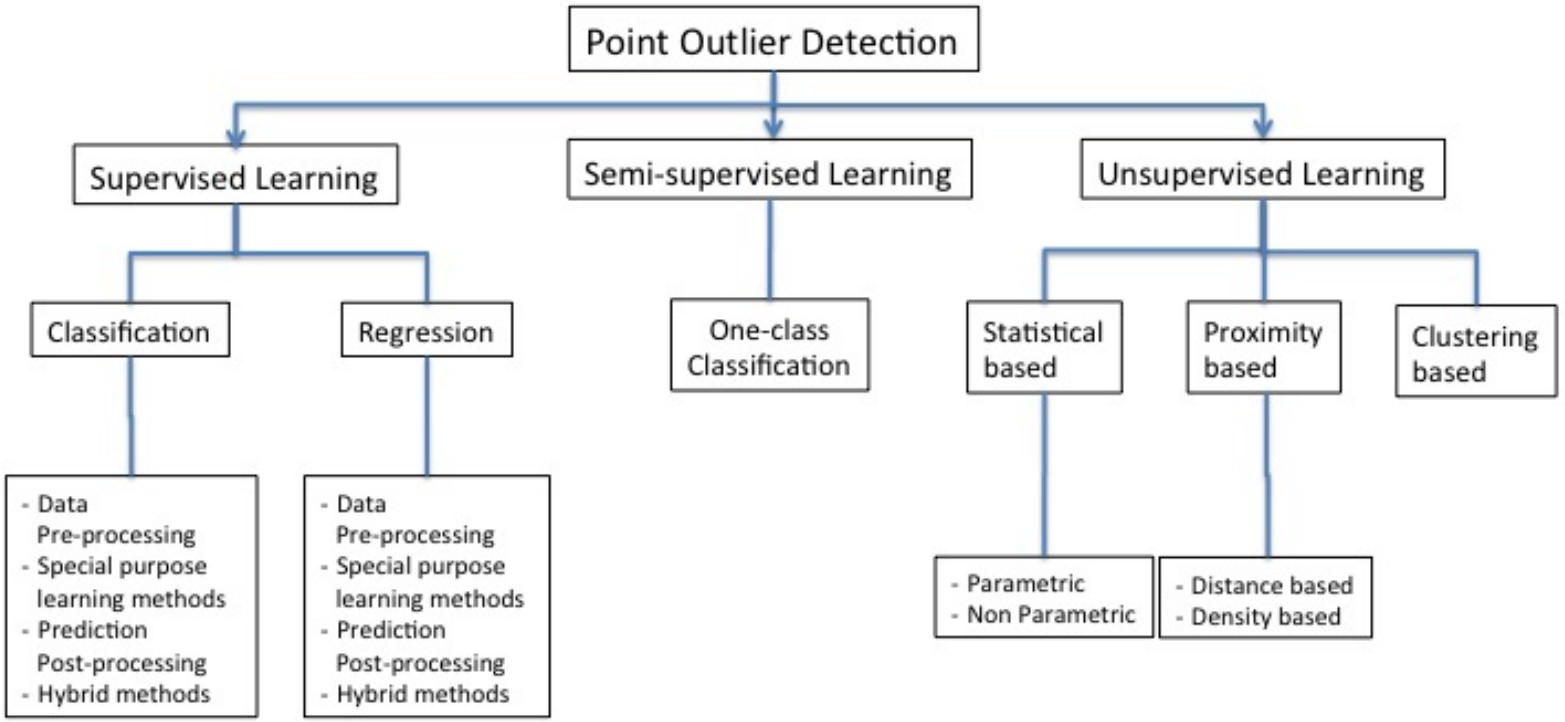
Learning Task: Unsupervised Outlier Detection

- Data set has no information on the behaviour of each instance;
- Most common case in real-life applications;
- It assumes that instances with normal behaviour are far more similar to each other;



- Limitations:
  - Computationally expensive in the training or testing phase;
  - Methods should combine global and local analysis;
  - In high-dimensional space, the contrast in the distances might not be meaningful.

Taxonomy of Outlier Detection Methods



Inadequacy of Standard Performance Metrics

- Standard performance metrics (e.g. *accuracy*, *error rate*, *MSE*, *MAD*) assume that all instances are equally relevant for the model performance.
- These metrics would give a good performance estimation to a model that performs well on normal (frequent) cases and bad on outlier (rare) cases.

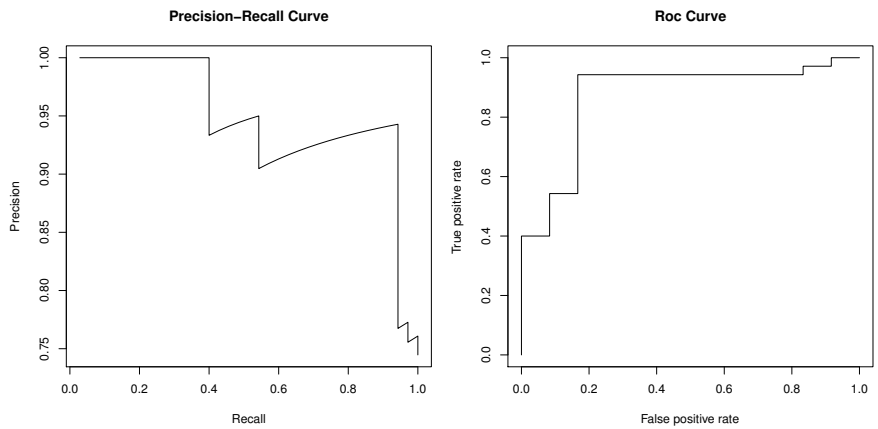
  Credit Card Fraud Detection:

  - data set D with only 1% of fraudulent transactions;
  - model M predicts all transactions as non-fraudulent;
  - M has a estimated accuracy of 99%;
  - yet, all the fraudulent transactions were missed!

Suitable Performance Metrics

Classification

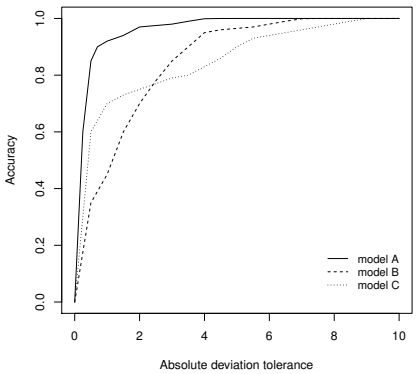
- metrics that focus on the performance on the minority (positive) class.
  - precision = TP/(TP+FP)
  - recall=TP/(TP+FN)
  - FPR = FP/(TN+FP)
  - F-measure: harmonic mean between precision and recall
  - AUC of ROC Curve: trade-off between recall and FPR
  - AUC of PR Curve: trade-off between precision and recall
  - ...

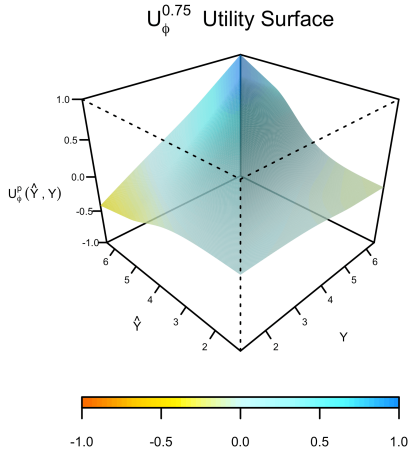


Suitable Performance Metrics (cont.)

Regression

- metrics with concepts inherited from classification that focus on the performance of a specific range of values
  - Utility Surfaces and Mean Utility
  - *precision*$^{\phi}$
  - *recall*$^{\phi}$
  - REC Curves
  - ...





See [Branco et al., 2016] for more details in performance metrics for imbalance predictive tasks

Outline

1 Outlier Detection Problem

2 Case Study: Detection of Potential Fraud Cases

3 Summary

Case Study: Detection of Potential Fraud Cases

- Data on items returning process from a large retail company:

Variable	Description
Supervisor	Code of the supervisor from the total of 887 supervisors.
Store	Code of the store from the total of 40 stores.
Region	Region of the store North, Center, South.
WeekDay	Type of day: Business days, Weekend or Holidays.
Period	Period of day : Morning, Afternoon, Night
ProdType	Product type: there are 18 different product types, such as Bakery, Fruits&Vegetables, Drink, Entertainment, etc.
TotalNr	Total number of items returned by the supervisor.
AvgVal	Average value of the items returned by the supervisor.

- 43206 transactions from December of 2014 to February of 2015.
- Objective: identify unusual cases that may suggest the occurrence of fraud in the returning of items process.

(Joint work with Ricardo Sousa, Eduarda Portela and João Gama)

Case Study: Detection of Potential Fraud Cases (cont.)

- The interpretability of the model that will indicate potential fraud cases for inspection is an important issue.

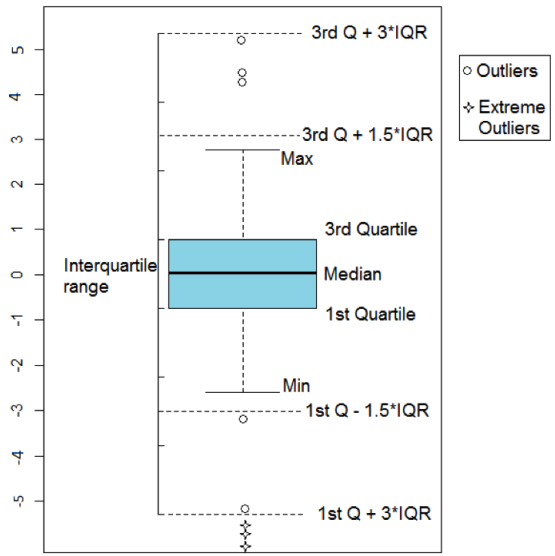
- An instance may be an outlier in a specific context (but not otherwise);
- Sometimes, there are contextual attributes that determine the context (sequential, spatial, etc) where we want outliers to be found.

- Our goal is two-fold:
  - detect the context where an outlier appears without *apriori* knowledge of contextual attributes;
  - assign a severity score to the outlier to make the inspection process more effective.

Case Study: Detection of Potential Fraud Cases (cont.)

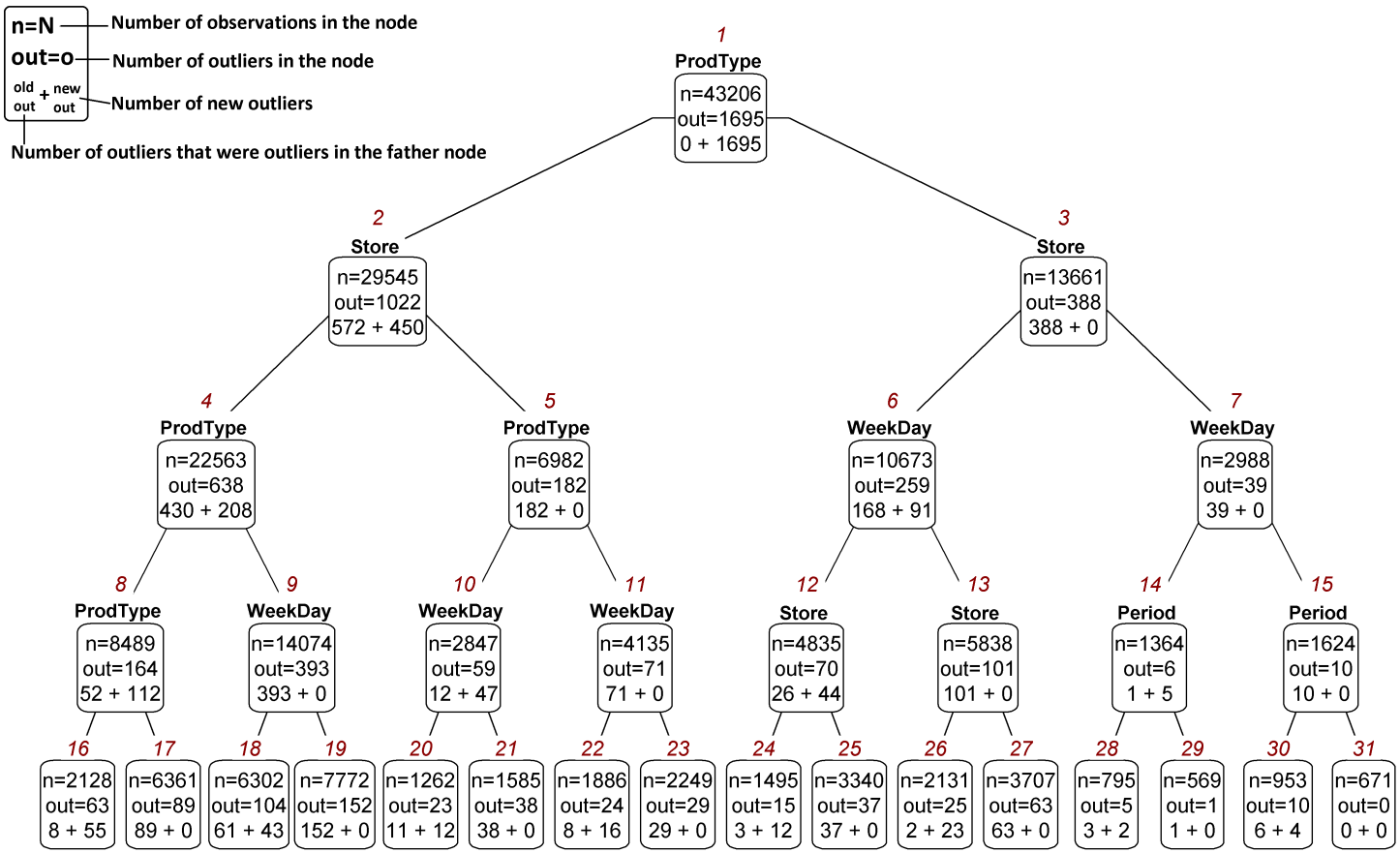
BoxplotTree

- We build a standard regression tree: it decreases the variance of the target variable on partitions formed by the attribute values
- At each partition, we identify the extreme outliers by the boxplot method.



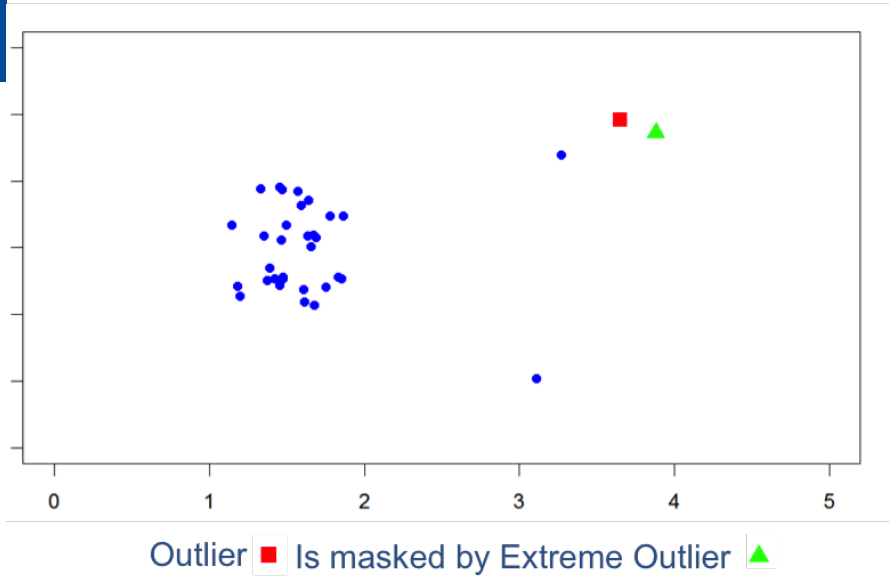
- Throughout the regression tree, on different nodes, some outliers appear while others disappear.

Case Study: Detection of Potential Fraud Cases (cont.)

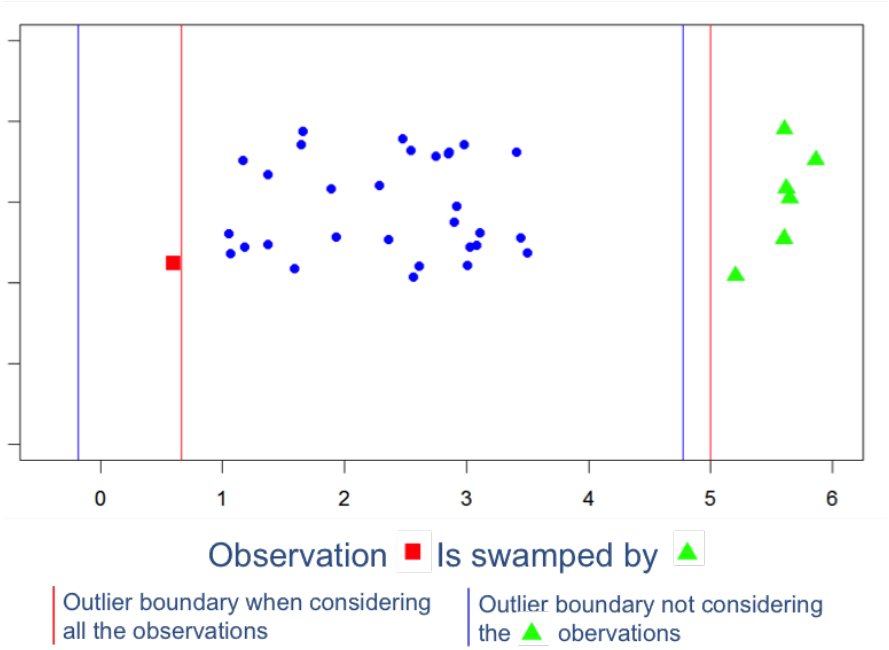


Case Study: Detection of Potential Fraud Cases (cont.)

Masking Effect



Swamping Effect



Simpson's Paradox

"(...) a trend that appears in different groups of data but disappears or reverses when these groups are combined."

Case Study: Detection of Potential Fraud Cases (cont.)

- Some of the suspicious cases found include:
  - In the products of "Beverages", "Hygiene" or "Pets Plants"
    - the average value of the returned items is around €9.37
    - in one identified case the returned item has a value around €173
  - In the products of "Fruits Vegetables" or "Bakery"
    - the average value of the returned items is around €2.55
    - in the two identified cases the return items value are over €35.
- It is important to notice that this method only suggests contextual outliers that may represent potential fraud cases.
- Only deeper inspection may confirm the reason for such deviant observations.

Case Study: Detection of Potential Fraud Cases (cont.)

- We have detected the context for each outlier observation.
- Now we want to obtain the outlinerness (severity) score for each identified observation.
- The goal is rank potential cases of fraud in order to make the inspection more effective.
- Our proposed score for each observation identified as an outlier at any node in the tree is based on:
  - the number of times it is identified as outlier across the tree;
  - the normalized distance to the median at each node where it appears;
  - the size of partition representing each node where it appears.

For more details see [Portela et al., 2019]

Outline

1 Outlier Detection Problem

2 Case Study: Detection of Potential Fraud Cases

3 Summary

Summary

- Outliers are not necessarily random noise.
- They can represent critical information that can trigger preventive or corrective actions.
- The interpretability of an outlier detection method is extremely important.
- The nature of the outlier detection problem is dependent on the application domain.
- Different approaches to this problem are necessary.
- Contextual and collective outliers are having increasing applicability in several real-world domains.
- There is much space for the development of new techniques in this area.

References

Aggarwal, C. (2013).

Outlier Analysis.

Springer New York.

Aggarwal, C. C. (2015).

Data Mining, The Texbook.

Springer.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016).

A survey of predictive modeling on imbalanced domains.

ACM Comput. Surv., 49(2):31:1–31:50.

Chandola, V., Banerjee, A., and Kumar, V. (2009).

Anomaly detection: A survey.

ACM Computing Surveys (CSUR), 41(3):15.

Hawkins, D. M. (1980).

Identification of Outliers.

Chapman and Hall.

References (cont.)

Hodge, V. J. and Austin, J. (2004).

A survey of outlier detection methodologies.

Artificial Intelligence Review, 22:2004.

Portela, E., Ribeiro, R. P., and Gama, J. (2019).

The search of conditional outliers.

Intelligent Data Analysis (IDA).

(to appear).

Weiss, G. M. (2004).

Mining with rarity: a unifying framework.

SIGKDD Explorations Newsletter, 6(1):7–19.