

---

# **Data Mining:**

## **Concepts and Techniques**

**(3<sup>rd</sup> ed.)**

**— Chapter 12 —**

Jiawei Han, Micheline Kamber, and Jian Pei  
University of Illinois at Urbana-Champaign &  
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

# Chapter 12. Outlier Analysis

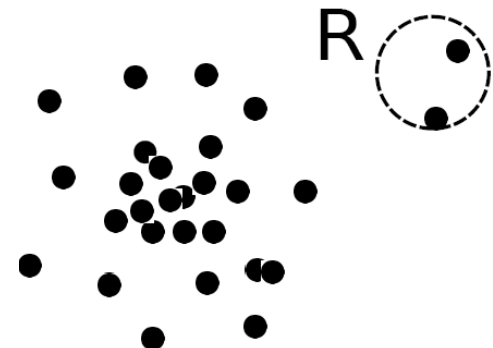
---

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Summary

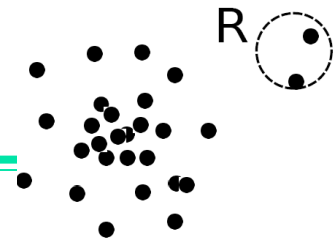


# What Are Outliers?

- **Outlier:** A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**
  - Ex.: Unusual credit card purchase, sports: Michael Jordon, Wayne Gretzky, ...
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: It violates the mechanism that generates the normal data
- Outlier detection vs. *novelty detection*: early stage, outlier; but later merged into the model
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis



# Types of Outliers (I)



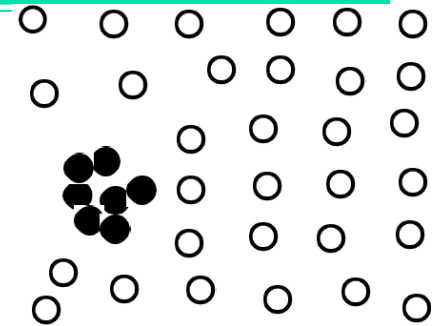
Global Outlier

- Three kinds: *global*, *contextual* and *collective* outliers
- **Global outlier** (or point anomaly)
  - Object is  $O_g$  if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation
- **Contextual outlier** (or *conditional outlier*)
  - Object is  $O_c$  if it deviates significantly based on a selected context
  - Ex. 80° F in Urbana: outlier? (depending on summer or winter?)
  - Attributes of data objects should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature
  - Can be viewed as a generalization of *local outliers*—whose density significantly deviates from its local area
  - Issue: How to define or formulate meaningful context?

# Types of Outliers (II)

## ■ Collective Outliers


- A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
- Applications: E.g., *intrusion detection*:
  - When a number of computers keep sending denial-of-service packages to each other
- Detection of collective outliers
  - Consider not only behavior of individual objects, but also that of groups of objects
  - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects.
- A data set may have multiple types of outlier
- One object may belong to more than one type of outlier



Collective Outlier

# Chapter 12. Outlier Analysis

---

- Outlier and Outlier Analysis
- Outlier Detection Methods 
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Summary

# Outlier Detection I: Supervised Methods

---

- Two ways to categorize outlier detection methods:
  - Based on whether user-labeled examples of outliers can be obtained:
    - Supervised, semi-supervised vs. unsupervised methods
  - Based on assumptions about normal data and outliers:
    - Statistical, proximity-based, and clustering-based methods
- **Outlier Detection I: Supervised Methods**
  - Modeling outlier detection as a classification problem
    - Samples examined by domain experts used for training & testing
  - Methods for Learning a classifier for outlier detection effectively:
    - Model normal objects & report those not matching the model as outliers, or
    - Model outliers and treat those not matching the model as normal
  - Challenges
    - Imbalanced classes, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers
    - Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)

# Outlier Detection II: Unsupervised Methods

---

- Assume the normal objects are somewhat “clustered” into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
- Ex. In some intrusion or virus detection, normal activities are diverse
  - Unsupervised methods may have a high false positive rate but still miss many real outliers.
  - Supervised methods can be more effective, e.g., identify attacking some key resources
- Many clustering methods can be adapted for unsupervised methods
  - Find clusters, then outliers: not belonging to any cluster
  - Problem 1: Hard to distinguish noise from outliers
  - Problem 2: Costly since first clustering: but far less outliers than normal objects
    - Newer methods: tackle outliers directly



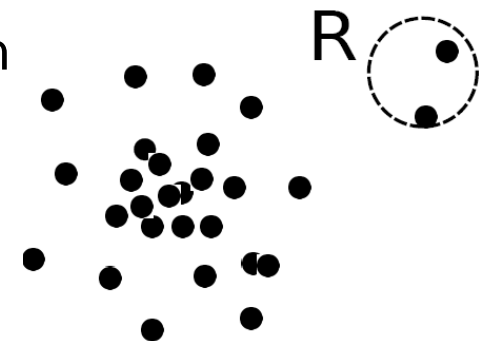
# Outlier Detection III: Semi-Supervised Methods

---

- Situation: In many applications, the number of labeled data is often small: Labels could be on outliers only, normal objects only, or both
- Semi-supervised outlier detection: Regarded as applications of semi-supervised learning
- If some labeled normal objects are available
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  - Those not fitting the model of normal objects are detected as outliers
- If only some labeled outliers are available, a small number of labeled outliers may not cover the possible outliers well
  - To improve the quality of outlier detection, one can get help from models for normal objects learned from unsupervised methods

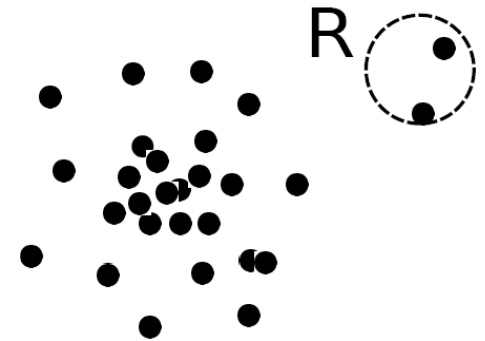
# Outlier Detection (1): Statistical Methods

- Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model (a stochastic model)
  - The data not following the model are outliers.
- Example (right figure): First use Gaussian distribution to model the normal data
  - For each object  $y$  in region  $R$ , estimate  $g_D(y)$ , the probability of  $y$  fits the Gaussian distribution
  - If  $g_D(y)$  is very low,  $y$  is unlikely generated by the Gaussian model, thus an outlier
- Effectiveness of statistical methods: highly depends on whether the assumption of statistical model holds in the real data
- There are rich alternatives to use various statistical models



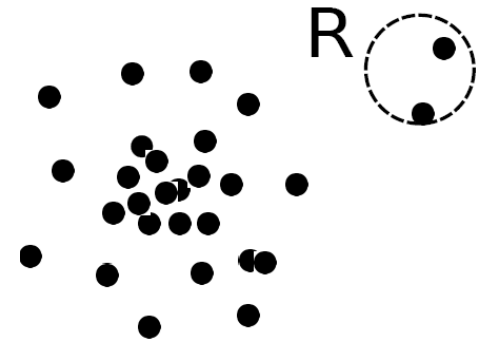
# Outlier Detection (2): Proximity-Based Methods

- An object is an outlier if the nearest neighbors of the object are far away, i.e., the **proximity** of the object is **significantly deviates** from the proximity of most of the other objects in the same data set
- Example (right figure): Model the proximity of an object using its 3 nearest neighbors
  - Objects in region R are substantially different from other objects in the data set.
  - Thus the objects in R are outliers
- The effectiveness of proximity-based methods highly relies on the proximity measure.
- In some applications, proximity or distance measures cannot be obtained easily.
- Often have a difficulty in finding a group of outliers which stay close to each other
- Two major types of proximity-based outlier detection
  - Distance-based vs. density-based




# Outlier Detection (3): Clustering-Based Methods

- Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters
- Example (right figure): two clusters
  - All points not in R form a large cluster
  - The two points in R form a tiny cluster, thus are outliers
- Since there are many clustering methods, there are many clustering-based outlier detection methods as well
- Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets



# Chapter 12. Outlier Analysis

---

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches 
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Summary

# Parametric Methods : Detection Univariate Outliers Based on Normal Distribution

- Univariate data: A data set involving only one attribute or variable
- Often assume that data are generated from a normal distribution, learn the parameters from the input data, and identify the points with low probability as outliers
- Ex: Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}
  - Use the maximum likelihood method to estimate  $\mu$  and  $\sigma$

$$\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^n \ln f(x_i | (\mu, \sigma^2)) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

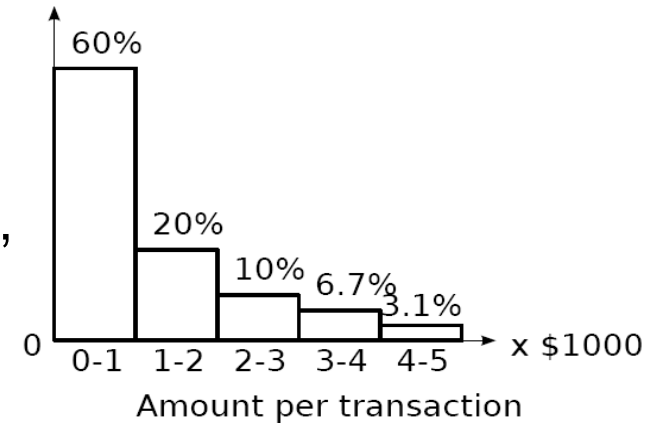
- Taking derivatives with respect to  $\mu$  and  $\sigma^2$ , we derive the following maximum likelihood estimates

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- For the above data with  $n = 10$ , we have  $\hat{\mu} = 28.61$   $\hat{\sigma} = \sqrt{2.29} = 1.51$
- Then  $(24 - 28.61) / 1.51 = -3.04 < -3$ , 24 is an outlier since  $\mu \pm 3\sigma$  region contains 99.7% data


# Non-Parametric Methods: Detection Using Histogram

- The model of normal data is learned from the input data without any *a priori* structure.
- Often makes fewer assumptions about the data, and thus can be applicable in more scenarios
- Outlier detection using histogram:
  - Figure shows the histogram of purchase amounts in transactions
  - A transaction in the amount of \$7,500 is an outlier, since only 0.2% transactions have an amount higher than \$5,000
- Problem: Hard to choose an appropriate bin size for histogram
  - Too small bin size → normal objects in empty/rare bins, false positive
  - Too big bin size → outliers in some frequent bins, false negative
- Solution: Adopt kernel density estimation to estimate the probability density distribution of the data. If the estimated density function is high, the object is likely normal. Otherwise, it is likely an outlier.



# Chapter 12. Outlier Analysis

---

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches 
- Clustering-Base Approaches
- Classification Approaches
- Summary




# Proximity-Based Approaches: Distance-Based vs. Density-Based Outlier Detection

---

- Intuition: Objects that are far away from the others are outliers
- Assumption of proximity-based approach: The proximity of an outlier deviates significantly from that of most of the others in the data set
- Two types of proximity-based outlier detection methods
  - Distance-based outlier detection: An object  $o$  is an outlier if its neighborhood does not have enough other points
  - Density-based outlier detection: An object  $o$  is an outlier if its density is relatively much lower than that of its neighbors

# Chapter 12. Outlier Analysis

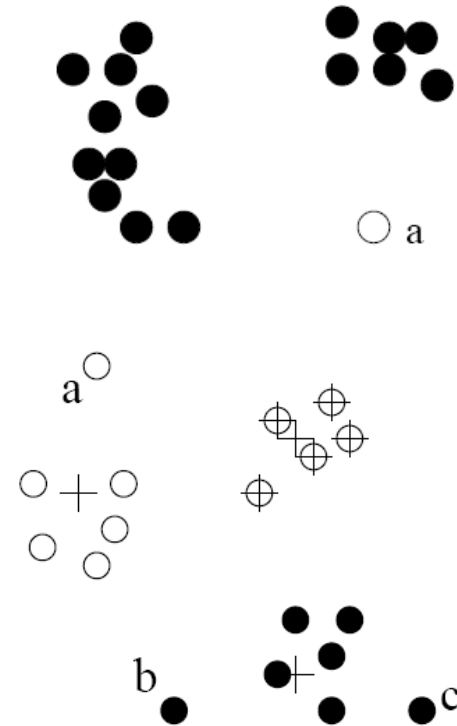
---

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches 
- Classification Approaches
- Summary

# Clustering-Based Outlier Detection (1 & 2):

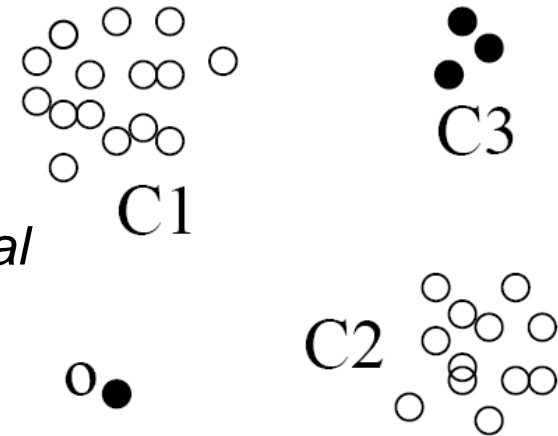
**Not belong to any cluster, or far from the closest one**

- An object is an outlier if (1) it does not belong to any cluster, (2) there is a large distance between the object and its closest cluster, or (3) it belongs to a small or sparse cluster
- Case 1: Not belong to any cluster
  - Identify animals not part of a flock: Using a density-based clustering method such as DBSCAN
- Case 2: Far from its closest cluster
  - Using k-means, partition data points of into clusters
  - For each object  $o$ , assign an outlier score based on its distance from its closest center
    - If  $\text{dist}(o, c_o)/\text{avg\_dist}(c_o)$  is large, likely an outlier
- Ex. Intrusion detection: Consider the similarity between data points and the clusters in a training data set
  - Use a training set to find patterns of “normal” data, e.g., frequent itemsets in each segment, and cluster similar connections into groups
  - Compare new data points with the clusters mined—Outliers are possible attacks



# Clustering-Based Outlier Detection (3): Detecting Outliers in Small Clusters

- *FindCBLOF*: Detect outliers in small clusters
  - Find clusters, and sort them in decreasing size
  - To each data point, assign a *cluster-based local outlier factor* (CBLOF):
    - If obj  $p$  belongs to a large cluster,  $CBLOF = \text{cluster\_size} \times \text{similarity between } p \text{ and cluster}$
    - If  $p$  belongs to a small one,  $CBLOF = \text{cluster size} \times \text{similarity betw. } p \text{ and the closest large cluster}$
- Ex. In the figure,  $o$  is outlier since its closest large cluster is  $C_1$ , but the similarity between  $o$  and  $C_1$  is small. For any point in  $C_3$ , its closest large cluster is  $C_2$  but its similarity from  $C_2$  is low, plus  $|C_3| = 3$  is small



# Clustering-Based Method: Strength and Weakness

---

## ■ Strength


- Detect outliers without requiring any labeled data
- Work for many types of data
- Clusters can be regarded as summaries of the data
- Once the cluster are obtained, need only compare any object against the clusters to determine whether it is an outlier (fast)

## ■ Weakness

- Effectiveness depends highly on the clustering method used—they may not be optimized for outlier detection
- High computational cost: Need to first find clusters
- A method to reduce the cost: Fixed-width clustering
  - A point is assigned to a cluster if the center of the cluster is within a pre-defined distance threshold from the point
  - If a point cannot be assigned to any existing cluster, a new cluster is created and the distance threshold may be learned from the training data under certain conditions

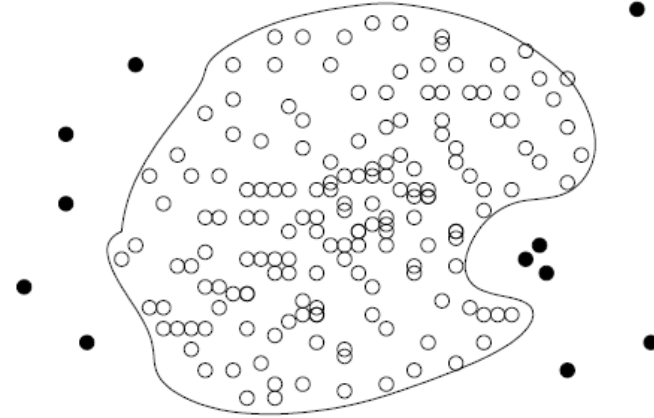
# Chapter 12. Outlier Analysis

---

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches 
- Summary

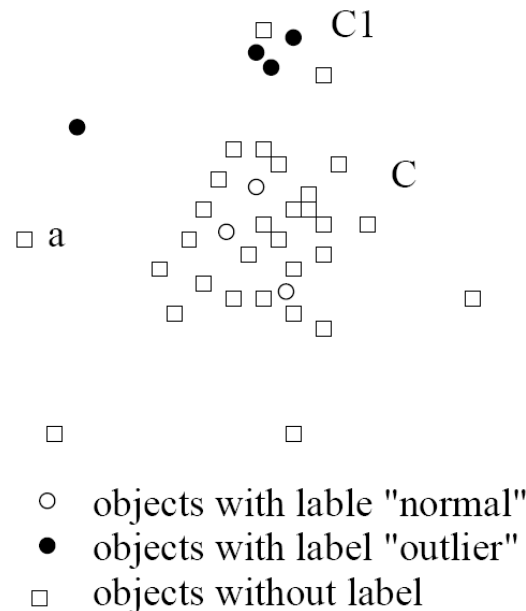
# Classification-Based Method I: One-Class Model

- Idea: Train a classification model that can distinguish “normal” data from outliers
- A brute-force approach: Consider a training set that contains samples labeled as “normal” and others labeled as “outlier”
  - But, the training set is typically heavily biased: # of “normal” samples likely far exceeds # of outlier samples
  - Cannot detect unseen anomaly
- One-class model: A classifier is built to describe only the normal class.
  - Learn the decision boundary of the normal class using classification methods such as SVM
  - Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers
  - Adv: can detect new outliers that may not appear close to any outlier objects in the training set
  - Extension: Normal objects may belong to multiple classes



# Classification-Based Method II: Semi-Supervised Learning

- Semi-supervised learning: Combining classification-based and clustering-based methods
- Method
  - Using a clustering-based approach, find a large cluster,  $C$ , and a small cluster,  $C_1$
  - Since some objects in  $C$  carry the label “normal”, treat all objects in  $C$  as normal
  - Use the one-class model of this cluster to identify normal objects in outlier detection
  - Since some objects in cluster  $C_1$  carry the label “outlier”, declare all objects in  $C_1$  as outliers
  - Any object that does not fall into the model for  $C$  (such as  $a$ ) is considered an outlier as well



- Comments on classification-based outlier detection methods
  - Strength: Outlier detection is fast
  - Bottleneck: Quality heavily depends on the availability and quality of the training set, but often difficult to obtain representative and high-quality training data



# Chapter 12. Outlier Analysis

---

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Base Approaches
- Clustering-Base Approaches
- Classification Approaches
- Summary



# Summary

---

- Types of outliers
  - global, contextual & collective outliers
- Outlier detection
  - supervised, semi-supervised, or unsupervised
- Statistical (or model-based) approaches
- Proximity-base approaches
- Clustering-base approaches
- Classification approaches
- Mining contextual and collective outliers
- Outlier detection in high dimensional data

# References (I)

---

- B. Abraham and G.E.P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66:229–248, 1979.
- M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.*, 10:521–538, 2006.
- F. J. Anscombe and I. Guttman. Rejection of outliers. *Technometrics*, 2:123–147, 1960.
- D. Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowl. Inf. Syst.*, 11:29–44, 2006.
- F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *TKDE*, 2005.
- C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. *SIGMOD’01*
- R.J. Beckman and R.D. Cook. Outlier...s. *Technometrics*, 25:119–149, 1983.
- I. Ben-Gal. Outlier detection. In *Maimon O. and Rockach L. (eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic, 2005.
- M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. *SIGMOD’00*
- D. Barbar’a, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia. Bootstrapping a data mining intrusion detection system. *SAC’03*
- Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris. A comparative study for outlier detection techniques in data mining. *IEEE Conf. on Cybernetics and Intelligent Systems*, 2006.
- S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. *KDD’03*
- D. Barbara, N. Wu, and S. Jajodia. Detecting novel network intrusion using bayesian estimators. *SDM’01*
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58, 2009.
- D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. In *CEC’02*

# References (2)

---

- E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Proc. 2002 Int. Conf. of Data Mining for Security Applications*, 2002.
- E. Eskin. Anomaly detection over noisy data using learned probability distributions. *ICML'00*
- T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1997.
- V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22:85–126, 2004.
- D. M. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.
- Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recogn. Lett.*, 24, June, 2003.
- W. Jin, K. H. Tung, and J. Han. Mining top-n local outliers in large databases. *KDD'01*
- W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. *PAKDD'06*
- E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. *KDD'97*
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. *VLDB'98*
- E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *VLDB J.*, 8:237–253, 2000.
- H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. *KDD'08*
- M. Markou and S. Singh. Novelty detection: A review—part 1: Statistical approaches. *Signal Process.*, 83:2481–2497, 2003.
- M. Markou and S. Singh. Novelty detection: A review—part 2: Neural network based approaches. *Signal Process.*, 83:2499–2521, 2003.
- C. C. Noble and D. J. Cook. Graph-based anomaly detection. *KDD'03*

# References (3)

---

- S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. *ICDE'03*
- A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.*, 51, 2007.
- X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Trans. on Knowl. and Data Eng.*, 19, 2007.
- Y. Tao, X. Xiao, and S. Zhou. Mining distance-based outliers from large databases in any metric space. *KDD'06*
- N. Ye and Q. Chen. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 17:105–112, 2001.
- B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. *ICDE'00*