Outlier Detection

Chapter 12 of Data Mining: Concepts and Techniques

JIAWEI HAN, MICHELINE KAMBER, JIAN PEI

PRESENTED BY: SHERRY ZHU EECS6412 WINTER 2017

MARCH 15, 2017

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

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 transaction
- Outliers ≠ Noise data
- 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
  Medical analysis

Types of Outliers

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
- 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
- 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*
- Issue: How to define or formulate meaningful context?

Types of Outliers (Cont'd)

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:
- 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

Challenges of Outlier Detection

- Modeling normal objects and outliers properly
- Application-specific outlier detection
- Handling noise in outlier detection
- Understandability

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods 
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

Outlier Detection 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
  - 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
  - Procedure: Find clusters, then outliers → those 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

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** thus is 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 many 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 IV: 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: First use **Gaussian distribution** to model the normal data (*on bb*)
  - 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
  - E.g., parametric vs. non-parametric

Outlier Detection V: 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 deviated** from the **proximity of most of the other object**s in the same data set

- Example: Model the proximity of an object using its 3 nearest neighbors(*on bb*)
  - Objects in region R are **substantially different from** other objects in the data set → 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 VI: 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: two clusters (*on bb*)
  - 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

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

Statistical Approaches

- Statistical approaches assume that the objects in a data set are generated by **a stochastic process** (idea)
- Methods are divided into two categories:
  parametric vs. *non-parametric*
- **Parametric method**
  - Assumes that the **normal data is generated by a parametric distribution** with parameter θ
  - The probability density function of the parametric distribution $f(x, \vartheta)$ gives the probability that object x is generated by the distribution
  - The **smaller** this value, the more likely x is an **outlier**

Statistical Approaches Cont'd

- **Non-parametric method**
  - **Not assume** an a-priori statistical model and determine the model from the input data
  - **Not completely parameter free** but consider the number and nature of the parameters are flexible and not fixed in advance
  - Examples: histogram and kernel density estimation

Parametric Methods I: 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 μ and σ

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

Parametric Methods I: Detection Univariate Outliers Based on *Normal Distribution* Cont'd

Taking derivatives with respect to μ and σ^2, we derive the following maximum likelihood estimates

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

For the above data with n = 10, we have

Then (24 – 28.61) /1.51 = – 3.04 < –3, 24 is an outlier since

$$\mu \pm 3\sigma \text{ region contains } 99.7\% \text{ data}$$

Parametric Methods I: The Grubb's Test

- Univariate outlier detection: The Grubb's test (maximum normed residual test) — **another statistical method** under normal distribution
  - For each object x in a data set, compute its z-score: x is an outlier if

$$z \geq \frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

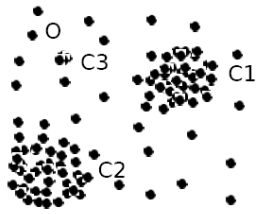
where $t^2_{\alpha/(2N),N-2}$ is the value taken by a t-distribution at a significance level of α/(2N), and N is the # of objects in the data set

Parametric Methods II: Detection of Multivariate Outliers

- Multivariate data: A data set involving **two or more attributes or variables**
- **Transform** the multivariate outlier detection task into a univariate outlier detection problem
- Method 1. **Compute Mahalaobis distance**
  - Let ō be the mean vector for a multivariate data set. Mahalaobis distance for an object o to ō is $MDist(o, \bar{o}) = (o - \bar{o})^T S^{-1} (o - \bar{o})$ where S is the covariance matrix
  - Use the Grubb's test on this measure to detect outliers
- Method 2. **Use χ^2 –statistic**: $\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$
  - where E_i is the mean of the *i*-dimension among all objects, and n is the dimensionality
  - If χ^2 –statistic is large, then object o_i is an outlier

Parametric Methods III: Using Mixture of Parametric Distributions

- Assuming data generated by a normal distribution could be sometimes overly simplified
- Example: The objects between the two clusters cannot be captured as outliers since they are close to the estimated mean



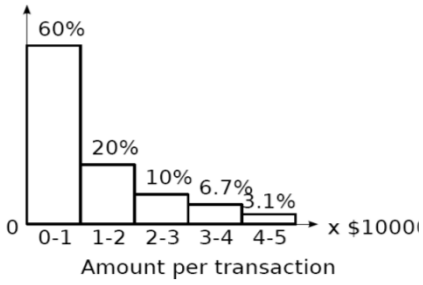
Parametric Methods III: Using Mixture of Parametric Distributions Cont'd

- To overcome this problem, assume the normal data is generated by **two normal distributions**. For any object o in the data set, the probability that o is generated by the **mixture of the two distributions** is given by

$$Pr(o|\Theta_1, \Theta_2) = f_{\Theta_1}(o) + f_{\Theta_2}(o)$$

  where $f_{\theta 1}$ and $f_{\theta 2}$ are the probability density functions of θ_1 and θ_2
- Then use EM algorithm to learn the parameters μ_1, σ_1, μ_2, σ_2 from data
- An object o is an outlier if it does not belong to any cluster

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:

Non-Parametric Methods: Detection Using Histogram Cont'd

- **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.

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

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

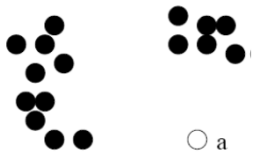
- Intuition: Objects that are **far away** from the others are outliers
- Assumption of proximity-based approach: The proximity of an outlier **deviates significantly** from that of most of the others in the data set
- Two types of proximity-based outlier detection methods
  - **Distance-based** outlier 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

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches 
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

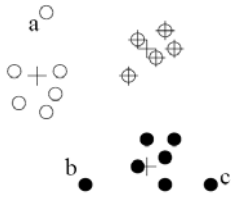
Clustering-Based Outlier Detection (1 & 2):
Not belong to any cluster, or far from the closest one

- 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 I: Not belong to any cluster
  - Identify animals not part of a flock: Using a density-based clustering method such as DBSCAN



Clustering-Based Outlier Detection (1 & 2):
Not belong to any cluster, or far from the closest one Cont'd

- 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 dist(o, c_o)/avg_dist(c_o) is large, likely an outlier



Clustering-Based Outlier Detection (1 & 2):
Not belong to any cluster, or far from the closest one Cont'd

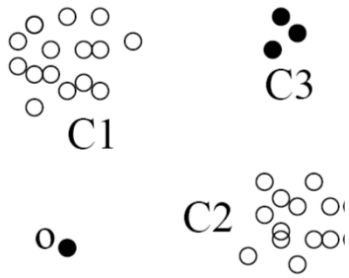
- Ex. *Intrusion detection*: Consider the similarity between data points and the clusters in a training data set
  - ➢ Use a training set to find **patterns of "normal"** data, e.g., frequent itemsets in each segment, and cluster similar connections into groups
  - ➢ **Compare** new data points with 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 = cluster_size X similarity between p and cluster
- If p belongs to a **small one**, CBLOF = cluster_size X similarity btw. p and the closest large cluster

Clustering-Based Outlier Detection (3): Detecting Outliers in Small Clusters Cont'd



- 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

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

Clustering-Based Method: Strength and Weakness Cont'd

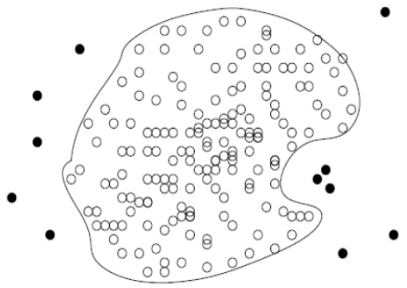
Weakness

1) Effectiveness depends **highly on the clustering method** used—they may **not be optimized** for outlier detection
2) **High computational cost**: Need to first find clusters
3) 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

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

Classification-Based Method I: One-Class Model



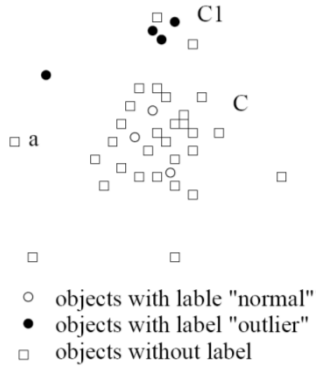
- **Idea**: Train a classification model that can distinguish "normal" data from outliers
- **A brute-force approach: Consider a training set that contains samples labeled as "normal" and others labeled as "outlier"**
  - But, the training set is typically heavily biased: # of "normal" samples likely far exceeds # of outlier samples
  - Cannot detect unseen anomaly

Classification-Based Method I: One-Class Model Cont'd

- 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



Classification-Based Method II: Semi-Supervised Learning Cont'd

Comments on Classification-based outlier detection methods:

Strength

- Outlier detection is efficient (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

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions

Summary

- Types of outliers (Do you remember?)
- Outlier detection (How about the detection methods?)
- **Statistical (or model-based) approaches**
- **Proximity-base approaches**
- **Clustering-base approaches**
- **Classification approaches**
- Mining contextual and collective outliers
- Outlier detection in high dimensional data

Agenda

- Outlier and Outlier Analysis
- Outlier Detection Methods
- Statistical Approaches
- Proximity-Based Approaches
- Clustering-Based Approaches
- Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary
- Discussions 

Discussions

•This "paper" is special, not a research paper → More of a survey for teaching purposes

• Open issues in Outlier and Anomaly Detection

- The application of Intrusion Detection Systems

1) Short of some fundamental redesign, today's outlier detection approaches in intrusion detection system will not be able to adequately protect tomorrow's networks against intrusions and attacks;
2) Inability to suppress false alarms;
3) No globally acceptable standard/ metric for evaluating an intrusion detection system

Questions?

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.

References (II)

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.

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 methdologies. *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*

References (III)

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*

S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. *ICDE'03*

References (IV)

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*