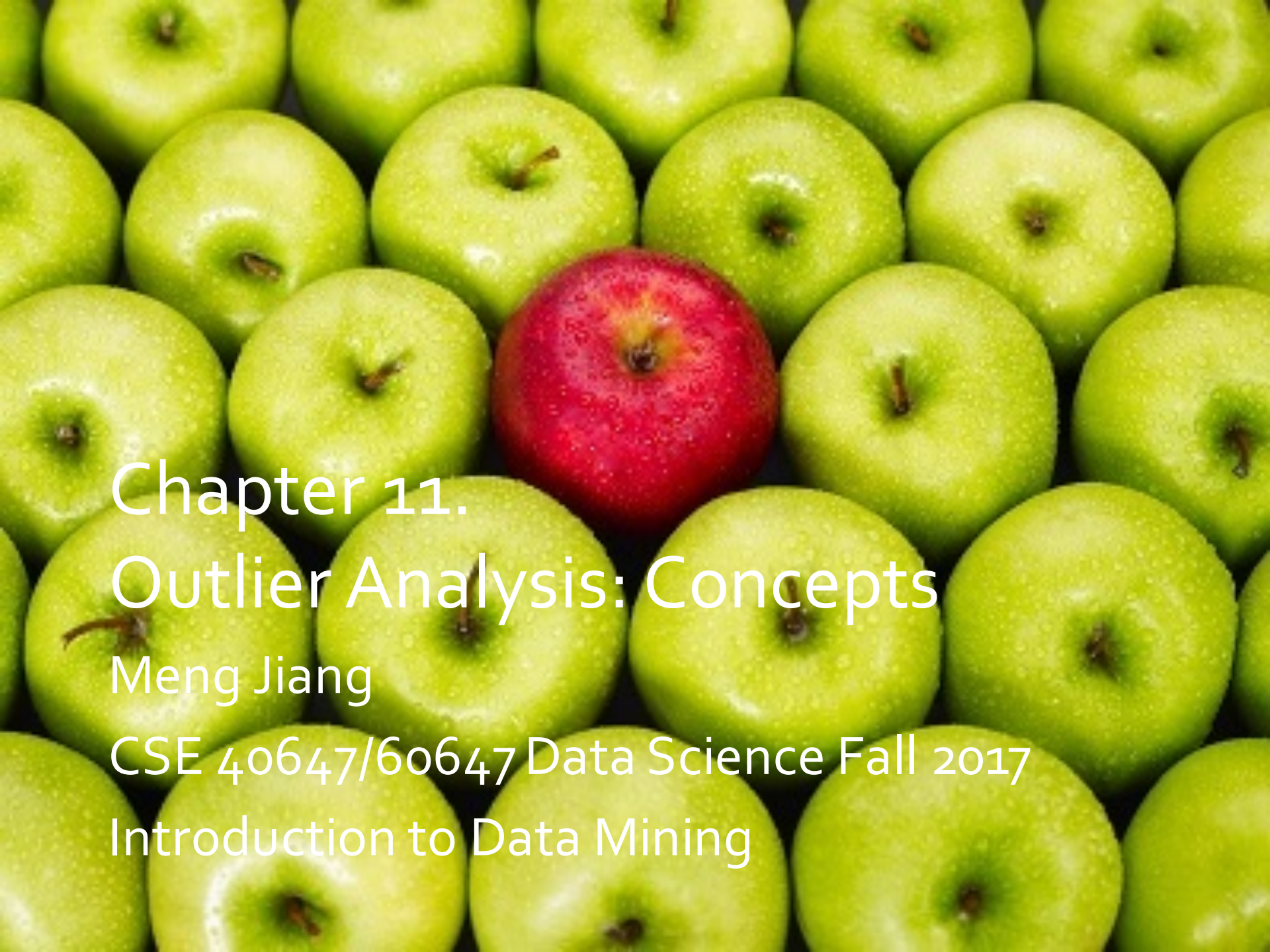


Chapter 11. Outlier Analysis: Concepts

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

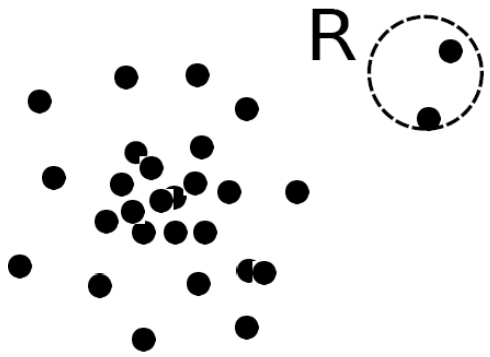
Introduction to Data Mining

Outlier Analysis

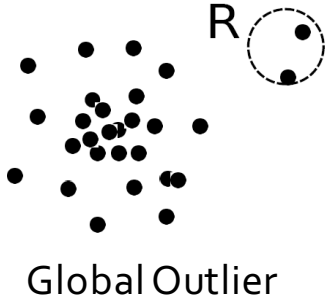
- **Basic Concepts**
- Outlier Detection Methods
- Statistical Approaches
- Clustering-Based Approaches
- Classification-Based Approaches

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 Jordan …
- 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
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis



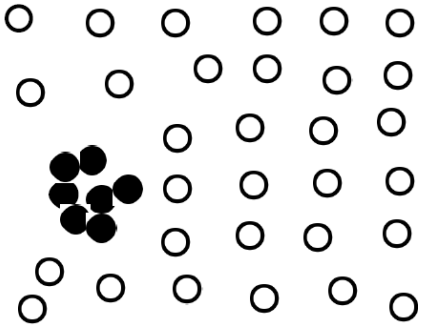
Types of Outliers (I)



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 Outlier

- 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

Challenges of Outlier Detection

- Modeling normal objects and outliers properly
  - Hard to enumerate all possible normal behaviors in an application
  - The border between normal and outlier objects is often a gray area
- Application-specific outlier detection
  - Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
- Handling noise in outlier detection
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection
- Understandability
  - Understand why these are outliers: Justification of the detection
  - Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism

Outlier Analysis

- Basic Concepts
- **Outlier Detection Methods**
- Statistical Approaches
- Clustering-Based Approaches
- Classification-Based Approaches

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

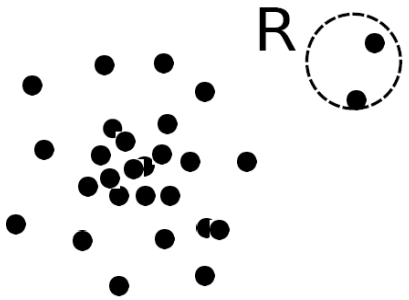
- Assume the normal objects are somewhat "clustered" into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
  - Normal objects may not share any strong patterns, but the 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 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

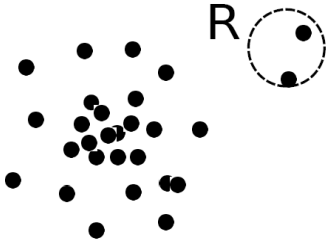
Outlier Detection (1): Statistical Methods

- 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: First use Gaussian distribution to model the normal data
  - For each object y in region R, estimate gD(y), the probability of y fits the Gaussian distribution
  - If gD(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 (2): 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

Outlier Analysis

- Basic Concepts
- Outlier Detection Methods
- **Statistical Approaches**
- Clustering-Base Approaches
- Classification Approaches

Statistical Approaches

- Statistical approaches assume that the objects in a data set are generated by a stochastic process (a generative model)
- Idea: learn a generative model fitting the given data set, and then identify the objects in low probability regions of the model as outliers
- 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
- 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

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

  - Taking derivatives with respect to μ and σ2, 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 $\hat{\mu} = 28.61 \;\; \hat{\sigma} = \sqrt{2.29} = 1.51$
  - Then (24 – 28.61) /1.51 = – 3.04 < –3, 24 is an outlier since

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

Parametric Methods II: 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 III: Detection of Multivariate Outliers

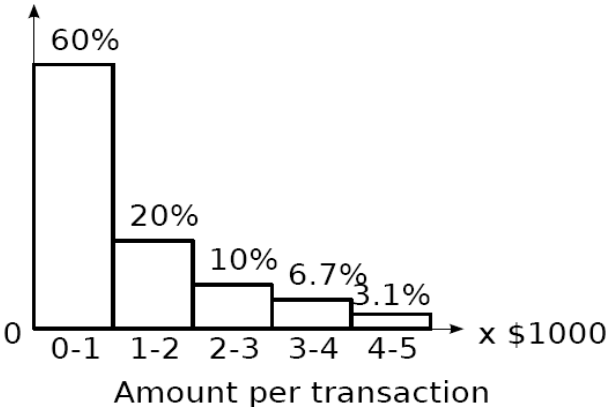
- Multivariate data: A data set involving two or more attributes or variables
- Transform the multivariate outlier detection task into a univariate outlier detection problem
- Method 1. Compute Mahalaobis distance
  - Let $\bar{o}$ be the mean vector for a multivariate data set. Mahalaobis distance for an object o to $\bar{o}$ 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

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.

References (1)

- B. Abraham and G.E.P. Box. Bayesian analysis of some outlier problems in time series. Biometrika, 1979.
- Malik Agyemang, Ken Barker, and Rada Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. Intell. Data Anal., 2006.
- Deepak Agarwal. Detecting anomalies in cross-classied streams: a bayesian approach. Knowl. Inf. Syst., 2006.
- C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. SIGMOD'01.
- M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Optics-of: Identifying local outliers. PKDD '99
- M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. SIGMOD'00.
- V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Comput. Surv., 2009.
- D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. Computational Intelligence, 2002.
- 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, 1997.
- R. Fujimaki, T. Yairi, and K. Machida. An approach to spacecraft anomaly detection problem using kernel feature space. KDD '05
- F. E. Grubbs. Procedures for detecting outlying observations in samples. Technometrics, 1969.

References (2)

- V. Hodge and J. Austin. A survey of outlier detection methodologies. Artif. Intell. Rev., 2004.
- Douglas M Hawkins. Identification of Outliers. Chapman and Hall, 1980.
- P. S. Horn, L. Feng, Y. Li, and A. J. Pesce. Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation. Clin Chem, 2001.
- W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. PAKDD'06
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98
- M. Markou and S. Singh.. Novelty detection: a review| part 1: statistical approaches. Signal Process., 83(12), 2003.
- M. Markou and S. Singh. Novelty detection: a review| part 2: neural network based approaches. Signal Process., 83(12), 2003.
- 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(12):3448{3470, 2007.
- W. Stefansky. Rejecting outliers in factorial designs. Technometrics, 14(2):469{479, 1972.
- X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. IEEE Trans. on Knowl. and Data Eng., 19(5):631{645, 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, 2001.