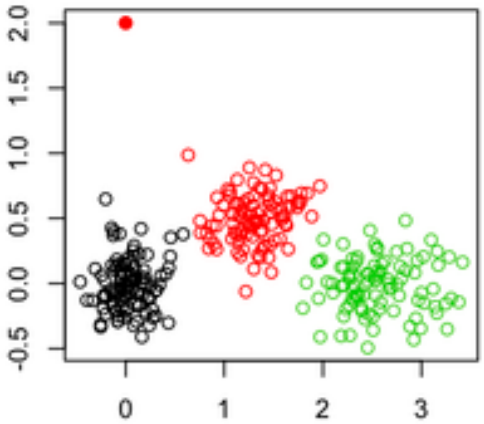
Outlier Detection

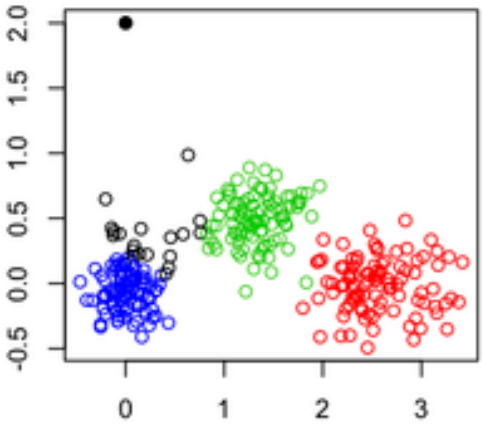
Motivation: Fraud Detection

http://i.imgur.com/ckkoAOp.gif

Techniques: Fraud Detection

- Features
- Dissimilarity
- Groups and noise





http://i.stack.imgur.com/tRDGU.png

Outlier Analysis

- "One person's noise is another person's signal"
- Outliers: the objects considerably dissimilar from the remainder of the data
  - Examples: credit card fraud, Michael Jordon, intrusions, etc
  - Applications: credit card fraud detection, telecom fraud detection, intrusion detection, customer segmentation, medical analysis, etc

Outliers and Noise

- Different from noise
  - Noise is random error or variance in a measured variable
- Outliers are interesting: an outlier violates the mechanism that generates the normal data
- Outlier detection vs. novelty detection
  - Early stage may be regarded as outliers
  - But later merged into the model

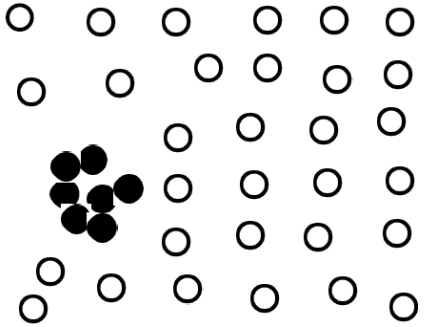
Types of Outliers

- Three kinds: global, contextual and collective outliers
  - A data set may have multiple types of outlier
  - One object may belong to more than one type of outlier
- Global outlier (or point anomaly)
  - An outlier object significantly deviates from the rest of the data set
- challenge: find an appropriate measurement of deviation

Contextual Outliers

- An outlier object deviates significantly based on a selected context
  - Ex. Is 10C in Vancouver an 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
- A generalization of local outliers—whose density significantly deviates from its local area
- Challenge: how to define or formulate meaningful context?

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
  - Application example: 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

Outlier Detection: Challenges

- 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
  - Example: clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations

Outlier Detection: Challenges

- Handling noise in outlier detection
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers
  - Noise 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 Detection Methods

- Whether user-labeled examples of outliers can be obtained
  - Supervised, semi-supervised, and unsupervised methods
- Assumptions about normal data and outliers
  - Statistical, proximity-based, and clustering-based methods

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)

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
- Many clustering methods can be adapted for unsupervised methods
  - Find clusters, then outliers: not belonging to any cluster

Unsupervised Methods: Challenges

- 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
- Challenges
  - Hard to distinguish noise from outliers
  - Costly since first clustering: but far less outliers than normal objects
- Newer methods: tackle outliers directly

Semi-Supervised Methods

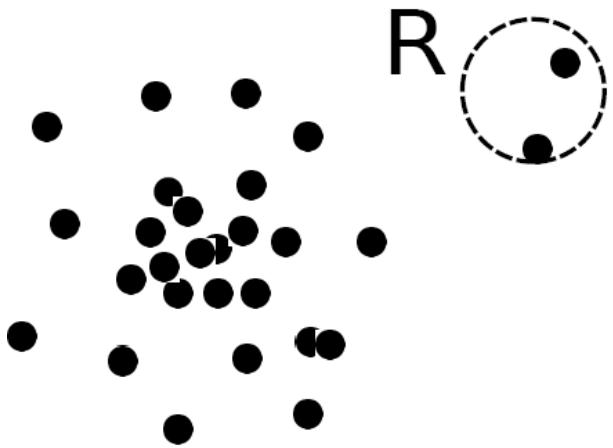
- In many applications, the number of labeled data is often small
  - Labels could be on outliers only, normal objects only, or both
- If some labeled normal objects are available
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  - Those not fitting 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

Pros and Cons

- 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
  - Parametric vs. non-parametric

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

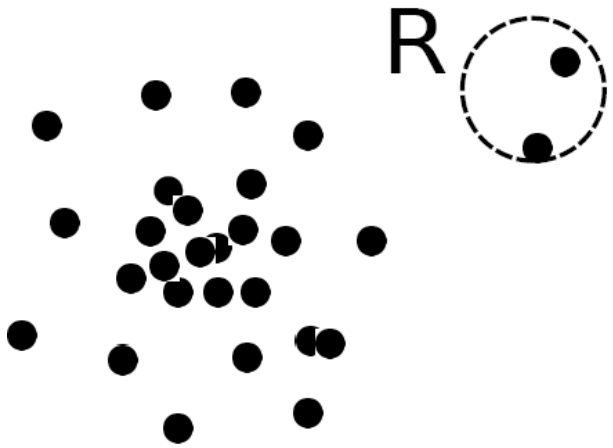


Pros and Cons

- 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 identifying a group of outliers that stay close to each other
- Two major types of proximity-based outlier detection methods
  - Distance-based vs. density-based

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



Challenges

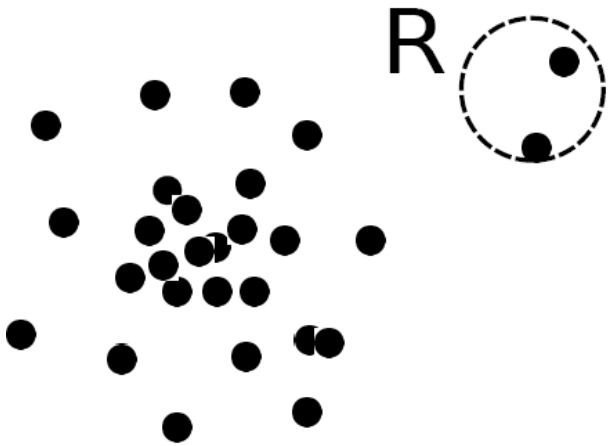
- Since there are many clustering methods, there are many clustering-based outlier detection methods as well
- Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets

Statistical Outlier Analysis

- Assumption: the objects in a data set are generated by a (stochastic) process (a generative model)
- Learn a generative model fitting the given data set, and then identify the objects in low probability regions of the model as outliers
- two categories: parametric versus non-parametric

Example

- Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model
  - The data not following the model are outliers.



Parametric Methods

- Assumption: the normal data is generated by a parametric distribution with parameter θ
- The probability density function of the parametric distribution $f(x \mid \theta)$ gives the probability that object x is generated by the distribution
- The smaller this value, the more likely x is an outlier

Univariate Outliers Based on Normal Distribution

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

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

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

Example

- Daily average temperature: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}
- Since n = 10, $\hat{\mu} = 28.61$ $\hat{\sigma} = \sqrt{2.29} = 1.51$
- Then (24 – 28.61) /1.51 = – 3.04 < –3, 24 is an outlier since μ ± 3σ contains 99.7% data

The Grubb's Test

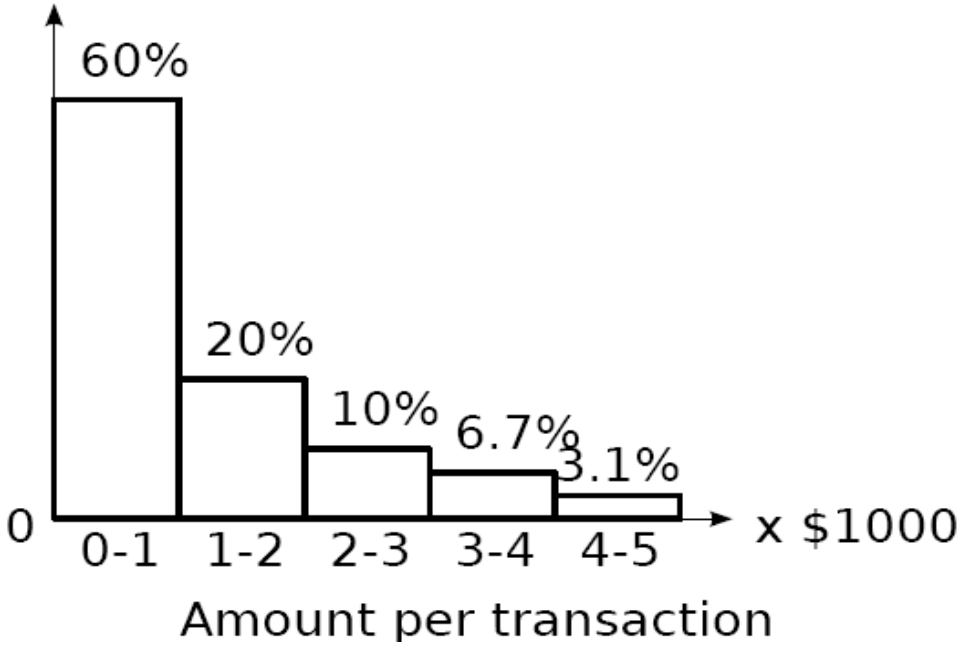
- Maximum normed residual test
- For each object x in a data set, compute its z-score
  - x is an outlier if $z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}}$
  - $t^2_{\frac{\alpha}{2N},N-2}$ is the value taken by a t-distribution at a significance level of α/(2N), and N is the number of objects in the data set

Non-parametric Method

- Not assume an a-priori statistical model, instead, 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

Histogram

- A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000



Challenges

- 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

To-Do List

- Read Chapters 12.1-12.3