C | A | U Christian-Albrechts-Universität zu Kiel Technische Fakultät

Inf-KDDM: Knowledge Discovery and Data Mining

Winter Term 2020/21

Lecture 7: Outlier Detection

Lectures: Prof. Dr. Matthias Renz

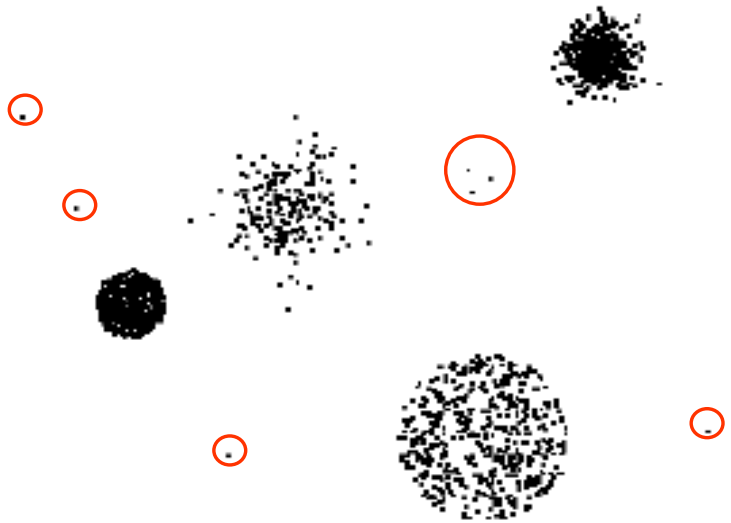
Exercises: Steffen Strohm

Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches

Outlier detection/ anomaly detection

- Goal: find objects that are considerably different from most other objects or unusual or in some way inconsistent with other objects
- Outliers / anomalous objects / exceptions
- Anomaly detection/ Outlier detection / Exception mining
- It is used either as a
  - Standalone task (anomalies are the focus)
  - Preprocessing task (to improve data quality)
- Applications
  - Fraud detection (credit card, telco)
  - Intrusion detection
  - Ecosystem disturbances
  - Public health
  - Medicine
  - Fault detection



Applications I

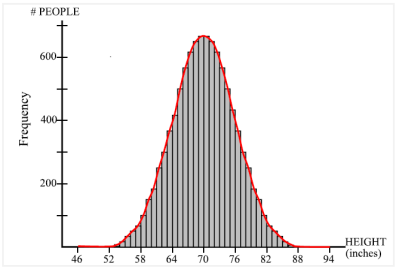
- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
- Public health
  - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
  - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Applications II

- Sports statistics
  - In many sports, various parameters are recorded for players in order to evaluate the players' performances
  - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
  - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
- Detecting measurement errors
  - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
  - Abnormal values could provide an indication of a measurement error
  - Removing such errors can be important in other data mining and data analysis tasks
  - "One person's noise could be another person's signal."
- ....

Causes of anomaly (few examples)

- Data from different classes
  - An object might be different from other objects because its of another class.
  - E.g. an attack connection in a network has different characteristics from a normal connection. Or, a person who commits credit card fraud belongs to a different class than persons using credit cards legally.
  - Such anomalies are the focus in Data Mining
- Natural variation
  - Many datasets can be modeled by statistical distributions e.g. Gaussian (most objects are near the center, s.t. the likelihood that an object differs significantly from this avg object is small).



  - e.g., an exceptional tiny elephant is not anomalous (not from another distribution), but in the sense of having an extreme size value.
- Data measurement and collection errors
  - erroneous measurements due to human/ measuring device errors, noise presence.
  - such errors should be eliminated since they just reduce the quality of data
- In practice, the techniques are not affected by the source of anomaly

What is an outlier?

- Definition of Hawkins [Hawkins 1980]:

> *"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"*

- Statistics-based intuition
  - Normal data objects follow a "generating mechanism", e.g. some given statistical process
  - Abnormal objects deviate from this generating mechanism

Discussion of the basic intuition based on Hawkins

- Data is usually multivariate, i.e., multi-dimensional
  - but, basic model is univariate, i.e., 1-dimensional
- There is usually more than one generating mechanism/statistical process underlying the "normal" data (comp. EM-Clustering)
  - but, basic model assumes only one "normal" generating mechanism
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
  - but, basic model assumes that outliers are rare observations
- A lot of models and approaches have evolved in the past years in order to exceed these assumptions

Variants of outlier detection problems

- Detect all anomalies in the database w.r.t. an anomaly threshold *t*: Given a database *D*, find all the data points $x \in D$ with anomaly scores $f(x)$ greater than some threshold *t*
- Detect top-*n* anomalies in the database: Given a database *D*, find all the data points $x \in D$ having the top-*n* largest anomaly scores *f(x)*
- Compute anomaly score for a query object: Given a database *D*, containing mostly normal (but unlabeled) data points, and a test point *x*, compute the anomaly score of *x* with respect to *D*

Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches

Basic application scenarios for outlier detection

Distinction based on the availability of class labels (for anomalies or normal instances)

- Supervised anomaly detection
  - In some applications, training data *with both normal and abnormal data* objects are provided
  - There may be multiple normal and/or abnormal classes
  - Often, the classification problem is *highly imbalanced*
- Semi-supervised anomaly detection
  - In some applications, only training data for the normal class(es) (or only the abnormal class(es)) are provided
- Unsupervised anomaly detection
  - In most applications there are no training data available
  - In such cases, the goal is to assign a score to each instance that reflects the degree to which the instance is anomalous.
  - This is the most common case.

Outlier detection vs clustering

- Are outliers just a side product of some clustering algorithms?
  - Many clustering algorithms do not assign all points to clusters but account for noise objects (e.g. DBSCAN, OPTICS, etc.)
  - Look for outliers by applying one of those algorithms and retrieve the noise set
- Problems
  - Clustering algorithms are optimized to find clusters rather than outliers
  - Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters (E.g. DBSCAN)
  - A set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers (E.g. OPTICS)
- So, outlier detection is a problem on its own.

Different classification approaches for outlier detection

- Global vs local outlier detection
  - Considers the set of reference objects relative to which each point's "outlierness" is judged
- Labeling vs scoring outliers
  - Considers the output of an algorithm
- Modeling properties
  - Considers the concepts based on which "outlierness" is modeled

- NOTE: we focus on models and methods for Euclidean data but many of those can be also used for other data types (because they only require a distance measure)

Global vs local outlier detection approaches

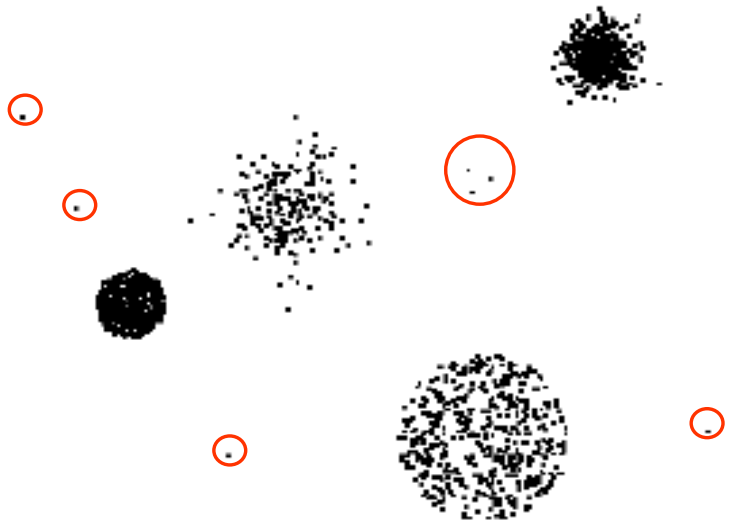
- Considers the resolution of the reference set w.r.t. which the “outlierness” of a particular data object is determined
- Global approaches
  - The reference set contains **all** other data objects
  - Basic assumption: there is only one normal mechanism
  - Basic problem: other outliers are also in the **reference set** and may falsify the results
- Local approaches
  - The reference contains a (small) **subset** of data objects
  - No assumption on the number of normal mechanisms
  - Basic problem: how to **choose** a proper reference set
- NOTE: Some approaches are somewhat in between (hybrid)
  - The resolution of the reference set varies e.g. from only a single object (local) to the entire database (global) automatically or by a user-defined input parameter

Labeling vs scoring outlier detection approaches

- Considers the output of an outlier detection algorithm
- Labeling approaches
  - Binary output
  - Data objects are labeled either as normal or outlier
- Scoring approaches
  - Continuous output
  - For each object an outlier score is computed (e.g. the probability of being an outlier)
  - Data objects can be sorted according to their scores
- Notes
  - Many scoring approaches focus on determining the top-*n* outliers (parameter *n* is usually given by the user)
  - Scoring approaches can be turned into labeling approaches if necessary (e.g. by defining a suitable threshold on the scoring values)

Outlier detection approaches w.r.t. modeling properties

- General steps
  - Build a profile of the “normal” behavior
    - i.e., patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile
- Types of anomaly detection schemes
  - Model-based (or, statistical approaches)
  - Distance-based
  - Density-based
  - Clustering-based



Outline

- Introduction
- Approaches for outlier detection
- Model-based or Statistical approaches
- Distance-based approaches
- Density-based approaches
- Clustering-based approaches