Chapter 11
Outlier Detection

11.1 Basic Concepts

Outline

- Basic concepts
  - What are outliers?
  - Types of outliers
  - Challenges in outlier detection
  - An overview of outlier detection methods
- Statistical approaches
- Proximity-based approaches
- Clustering and classification-based approaches

Motivation

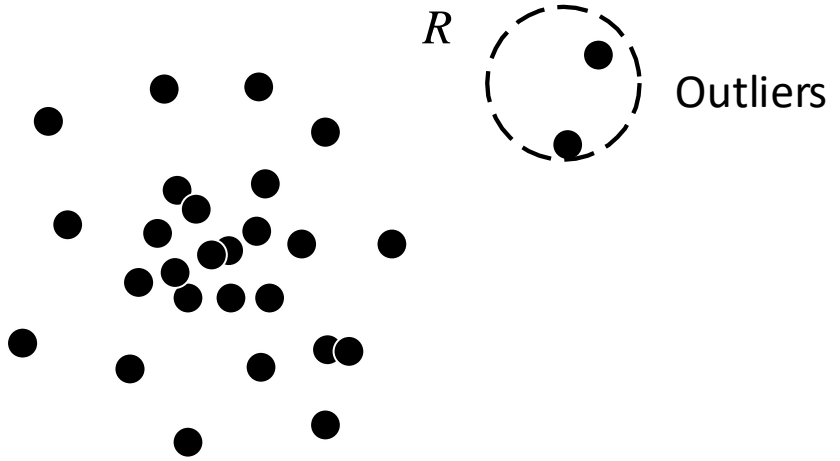
- How to detect suspicious credit card transactions, such as transactions of unusual amounts and multiple onsite transactions within 10 minutes in two locations hundreds of miles away?
- Most transactions are normal and only very few are anormal
- Outlier detection (also known as anomaly detection) is the process of finding outliers or anomalies, the data objects with behaviors that are very different from expectation
- Many applications, such as medical care, public safety and security, industry damage detection, image processing, sensor and video network surveillance, national security, and system intrusion detection

Outlier Detection versus Clustering

- Two highly related but different tasks
- Difference in purpose
  - Clustering finds the majority patterns in a data set and organizes the data accordingly
  - Outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns
- Difference in methodology
  - Outlier detection might also use supervision during the detection process
  - Clustering analysis is typically unsupervised in nature

11.1.1 What Are Outliers?

- Assume that a given statistical process is used to generate a set of data objects
- An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism



Outliers versus Noise

- Noise is a random error or variance in a measured variable, and thus is not interesting in data analysis, including outlier detection
  - Example: a bigger lunch one day or one more cup of coffee than usual
- Outliers are interesting because they are suspected of not being generated by the same mechanism as the rest of the data
  - It is important to justify why the detected outliers are generated by some other mechanisms
  - Example: make various assumptions on the rest of the data and show that the detected outliers violate those assumptions significantly

In a dog vs cat detector

Noise

Outlier

Outliers versus Noise (some ideas that can help)

Characteristic	Noise	Outlier
Consistency	Often affects a single feature	Affects multiple features consistently
Frequency	May appear randomly or sporadically	Often isolated but can indicate a pattern
Distance	Close but randomly scattered around normal data	Distant from central clusters or data points
Impact on Analysis	Minimal impact, adds variability	Can shift clusters, regression, or trends
Identification Technique	Smoothing, low Z-scores, DBSCAN noise label	High Z-scores, Isolation Forest, LOF

Outlier Detection and Novelty detection

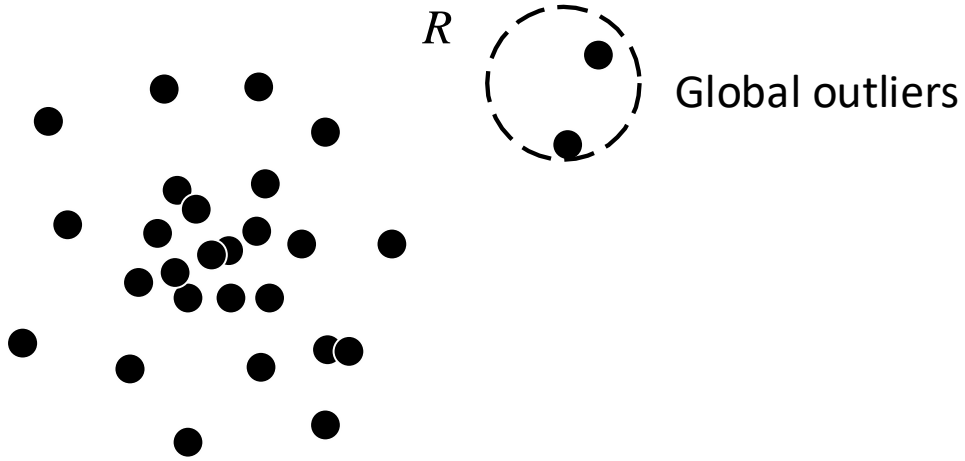
- Example: monitoring a social media web site where new content is incoming, novelty detection may identify new topics and trends in a timely manner
  - Novel topics may initially appear as outliers
- To this extent, outlier detection and novelty detection share some similarity in modeling and detection methods.
- Difference: in novelty detection, once new topics are confirmed, they are usually incorporated into the model of normal behavior so that follow-up instances are not treated as outliers anymore

11.1.2 Types of Outliers

- Global outliers
- Contextual (or conditional) outliers
- Collective outliers

Global Outliers

- A data object is a global outlier (or point anomaly) if it deviates significantly from the rest of the data set
  - Global outliers are the simplest type of outliers
- Most outlier detection methods are aimed at finding global outliers



Detecting Global Outliers

- It is critical to find an appropriate measurement of deviation with respect to the application in question
- Global outlier detection is important in many applications
  - Example: intrusion detection in computer networks
  - Example: trading transaction auditing systems

Contextual Outliers

- In a given data set, a data object is a contextual (or conditional) outlier if it deviates significantly with respect to a specific context of the object
- Example: “It is 28 degree Celsius today” – Is it exceptional and thus an outlier?
  - Yes, if today is in winter in Toronto, but no if today is in summer in Toronto
- In contextual outlier detection, the context has to be specified as part of the problem definition

Contextual versus Behavioral Attributes

In contextual outlier detection, the attributes of the data objects in question are divided into two groups: Contextual and Behavioral

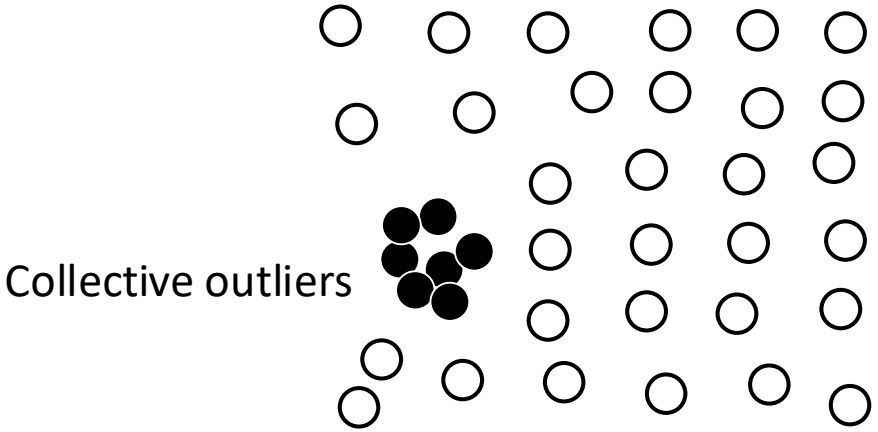
- The contextual attributes of a data object define the object's context
  - In the temperature example, the contextual attributes may be date and location
- The behavioral attributes define the object's characteristics, and are used to evaluate whether the object is an outlier in the context to which it belongs
  - In the temperature example, the behavioral attributes may be the temperature, humidity, and pressure
- Whether a data object is a contextual outlier depends on not only the behavioral attributes but also the contextual attributes

Contextual Outliers versus Global and Local Outliers

- Contextual outliers are a generalization of local outliers
  - An object in a data set is a local outlier if its density significantly deviates from the local area in which it occurs
- Global outlier detection can be regarded as a special case of contextual outlier detection where the set of contextual attributes is empty – global outlier detection uses the whole data set as the context
- Contextual outlier analysis provides flexibility to users in that one can examine outliers in different contexts, which can be highly desirable in many applications
- The quality of contextual outlier detection in an application depends on the meaningfulness of the contextual attributes, in addition to the measurement of the deviation of an object to the majority in the space of behavioral attributes

Collective Outliers

- Given a data set, a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set, while importantly the individual data objects may not be outliers
- Example: a student sick case in a class may not be considered outlying, but it is outlying if 10% of students are sick on a single day



Applications of Collective Outliers

- In intrusion detection, a denial-of-service package from one computer to another is considered normal, and not an outlier at all
  - However, if several computers keep sending denial-of-service packages to each other, they as a whole should be considered as a collective outlier
  - The computers involved may be suspected of being compromised by an attack
- A stock transaction between two parties is considered normal
  - However, a large set of transactions of the same stock between two or a small number of parties in a short period are collective outliers because they may be evidence of some people manipulating the market
- In collective outlier detection we have to consider not only the behavior of individual objects, but also that of groups of objects
- To detect collective outliers, we need background knowledge of the relationship among data objects, such as distance or similarity measurements between objects

Comparison among Multiple Types of Outliers

- A data set can have multiple types of outliers
- An object may belong to more than one type of outlier
- Different outliers may be used in various applications or for different purposes
- Global outlier detection is the simplest
- Context outlier detection requires background information to determine contextual attributes and contexts
- Collective outlier detection requires background information to model the relationship among objects to find groups of outliers

11.1.3 Challenges of Outlier Detection

- Modeling normal objects and outliers effectively
  - The border between data normality and abnormality (outliers) is often not clear-cut
- Application-specific outlier detection
  - The relationship among objects highly depends on applications
- Handling noise in outlier detection
  - Noise often unavoidably exists in data collected in many applications
  - Low data quality and the presence of noise bring a huge challenge to outlier detection
- Interpretability: a user may want to not only detect outliers, but also understand why the detected objects are outliers

11.1.4 Overview of outlier detection methods

- Categories Based on Data Labels
  - Supervised
  - Semi-Supervised
  - Unsupervised Methods
- Categories Based on Assumptions about Outliers vs. Normal Data
  - Statistical methods
  - Proximity-based methods
  - Reconstruction-based methods

Supervised, Semi-Supervised, and Unsupervised Methods

- Supervised methods model data normality and abnormality
  - Imbalanced classes (fewer outliers than normal data).
  - **Example Scenario**: Credit Card Fraud Detection
- Unsupervised outlier detection methods make an implicit assumption: the normal objects are somewhat “clustered”
  - **Example Scenario**: Network Intrusion Detection
- Semi-supervised methods: although obtaining some labeled examples is feasible, the number of such labeled examples is often small
  - **Example Scenario**: Industrial Equipment Failure Prediction

Statistical Methods, Proximity-based Methods, and Reconstruction-based Methods

- Statistical methods (also known as model-based methods) make assumptions of data normality
  - Statistical outlier detection assumes normal data follows a probabilistic model.
  - Example Scenario: Medical Diagnosis
- Proximity-based methods assume that an object is an outlier if the nearest neighbors of the object are far away in feature space
  - Example Scenario: Customer Behavior Analysis in Retail
- Reconstruction-based methods: matrix-factorization based methods and pattern-based compression methods
  - The normal data samples often share certain similarities, they can often be represented in a more succinct way, compared with their original representation
  - With the succinct representation, we can well reconstruct the original representation of the normal samples.
  - Example Scenario: Defective Product Detection in Manufacturing