**Data Mining Chapter 12**

**Outliers**==**anomalies**==**abnormal data**

= data object that deviates significantly from the rest of the objects (as though generated from dif. mechanism)

Unlike **noisy data**== random error in a measured variable (not actual outliers but variation in data) - noise should be removed before outlier detection

**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

***novelty detection vs outliers:*** critical difference is that in novelty detection, once differing values are confirmed, they are incorporated into the model of normal behavior so that follow-up instances are not treated as outliers anymore

* “included outliers”

**Types of outliers:**

**Global outliers:** Data object that deviates significantly from the rest of the data set

Critical: appropriate measure of deviation

**Contextual outlier==conditional outlier** deviates significantly with respect to a specific context of the object (conditional to on context)

e.g. temperature of 28C is an outlier in winter but not in summer

* + Contextual attributes: define the context (date & location)
  + Behavioral attributes: characteristics used to check if it’s an outlier (temperature)
* Whether an object is an outlier depends on both attributes

**Collective Outliers:** a subset of data objects can become an outlier as a whole even though the individual objects would not be considered outliers

**e.g.** one delivery with delay isn’t an outlier but 100 delayed deliveries in a day are a col. outlier

**Challenges of Outlier Detection**

**Modeling normal objects and outliers effectively**

* building a model for data normality is challenging
* Boarder between normality/anormality

**Application-specific outlier detection**

* outliers depend strongly on what is being measures -> no one rule to justify outliers

**Handling noise in outlier detection**

* Noise distorts and blurs outlier detection
* Noise can be accidentally identified as outlier and vice versa

**Understandability**

* Why are the detected objects outliers -> learn from them

**12.2 Outlier Detection Methods**

**Supervised**

* + Domain experts sample data
    - Label & model **normal objects** deviations = outliers
    - Label & model **outliers** deviations = normal
  + Challenges:
    - Classes (outliers/normal) are imbalances (more outliers than normal values) -> artificial outliers needed to make up for it

**Unsupervised**

**Assumption:**

* + normal objects are clustered -> normal objects follow a pattern more frequently than outliers
  + Not all values have to be part of the same pattern/group -> multiple groups but outliers are far from either one of them
  + Challenges:
    - If data is evenly distributed and outliers aren’t
    - Data objects identified as outliers might be noise

**Semi-supervised Methods**

* + Used where there’s only a few labeled examples for outliers/normal data
  + Like learning method: takes objects that are close to labeled normal data and uses these additional values to create a model of normal objects

**Statistical Methods**==**model based methods**

* assumption of data normality – assuming that data objects follow a model and data not following the model = outliers
* Highly dependent on whether the assumptions made for the model are true for the data

**Proximity-Based Methods**

* Assumption that objects are outliers when nearest neighbor objects are far away
* Highly dependent on proximity measure used
* Difficulty detecting outliers if they are close to each other

**Clustering-Based Methods**

**Normal data ->** large, dense clusters

**Outliers ->** small, sparse clusters or no cluster

**12.3 Statistical Approaches**

**Parametric methods:** assumption that normal data objects are generated by a parametric distribution with parameter probability density function gives prob. To an object -> the smaller the value the more likely is the object an outlier (parameters are fixed in advance) – **assumed model**

**Univariate outliers based on normal distribution**

**=** univariate data with only one attribute/variable

* + - Maximum likelihood
    - Outliers when over 3sd away from mean of estimated distribution
    - the region between *Q*1􀀀1.5\_*IQR* and *Q*3C1.5\_*IQR* contains 99.3% of the objects
    - boxplot

**multivariate outliers**

= two or more attributes/variables

* + - aim is to transform multivariate outlier detection into univariate outlier detection

**mixture of parametric distributions**

* + - assumption that data is generated by a **mixture** of parametric distributions
    - e.g. multiple normal distributions within clusters

**Non-parametric methods**: not a priori model but tries to determine model from input data (assumes that number/nature of parameters is flexible) – learned model from data

* + - **outlier detection with histogram**
      * **construct histogram, careful with bin size**
      * **outlier detection – if object falls into bin -> normal**

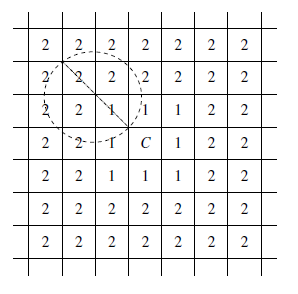
**12.4 Proximity-Based Approaches**

using distance measures to quantify similarity

**Assumption:** proximity of the outlier to the nearest neighbor significantly deviates from the prox. to most other objects in the data set

**Distance based**

* + - checks for each object within a set that should be tested for outliers and compares them to a neighborhood of the object with a set distance (r) if most of the object in the set are far from the neighborhood then the object can be regarded as an outlier

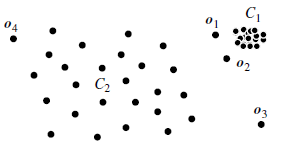
**Grid-Based Method**

* + - 1’s are level-1 cell property
    - Cells one or two cells away from C are level2 cells

Objects are organized into a group using a grid

Determine whether all objects in a cell are outliers or nonoutliers -> no need to check individually

**Density-Based Outlier Detection**

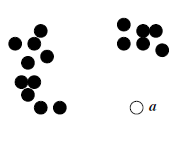
* + - Outlier is remote from the majority of the data
    - To detect distance we need two global parameters which are applied to every outlier object
    - O1 and o2 are not distance based outliers to C1 when using the distance measure in C2
    - but they can be outliers when considered locally to C1
    - o4 is not considered an outlier even though distance to neighbor is greater than o1 distance to neighbor but it is local to cluster C2 which is sparse and therefore the distance is not far in comparison

**12.5 Clustering-Based Approaches**

Detect outliers by examining the relationship between objects and clusters -> outlier belongs to small/remote cluster or no cluster

**Approaches:**

* + - if it doesn’t belong in a cluster -> outlier
    - large distance to closest cluster -> outlier



* + - part of small/sparse cluster -> all objects in cluster are outliers

**12.6 Classification-Based Approaches**

**General idea:** to train a classification model that can distinguish normal data from outliers

* Uses different models on the outlier object to determine whether it falls into the model or not
* Can incorporate human domain knowledge