S2-20_DSECLZC415: Data Mining
(Lecture #14 – Outlier Analysis)

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

- *The slides presented here are obtained from the authors of the books and from various other contributors. I hereby acknowledge all the contributors for their material and inputs.*
- *I have added and modified a few slides to suit the requirements of the course.*

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Data Mining

Outlier Analysis

What Are Outliers/Anomalies?

Outlier: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism**

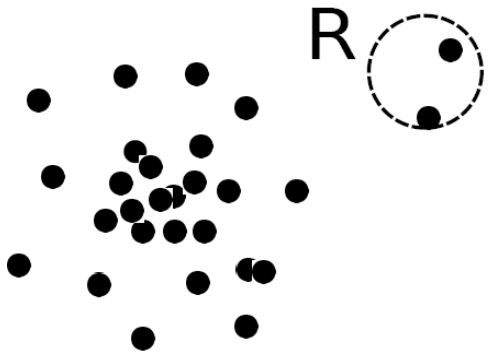
Outliers are different from the noise data

- 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



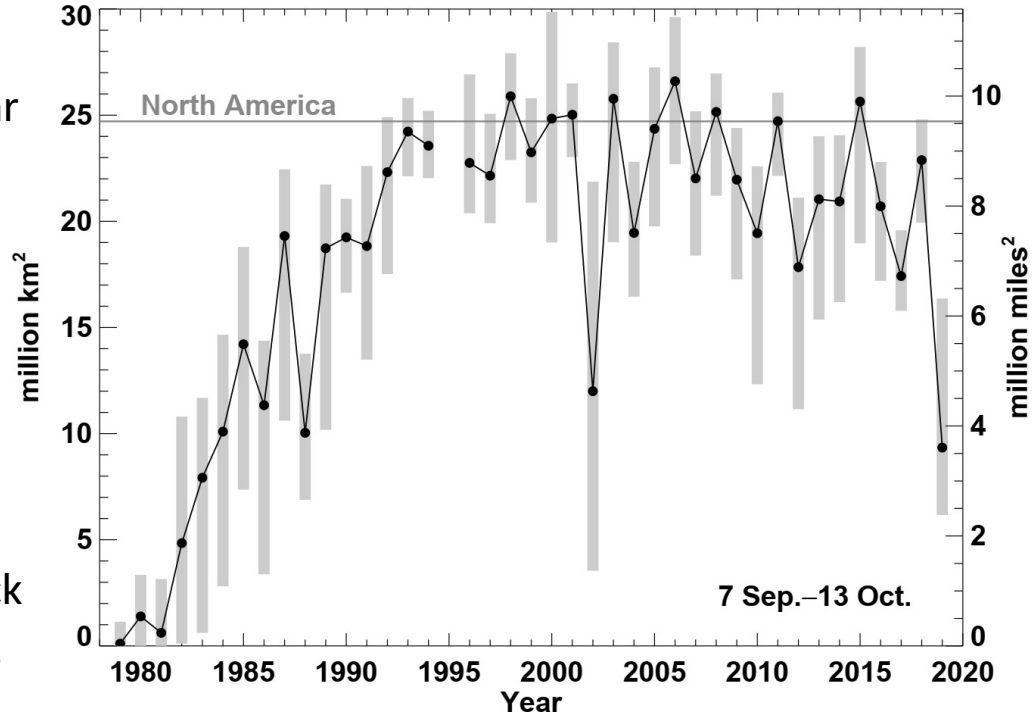
Importance of Anomaly Detection

Ozone Depletion History

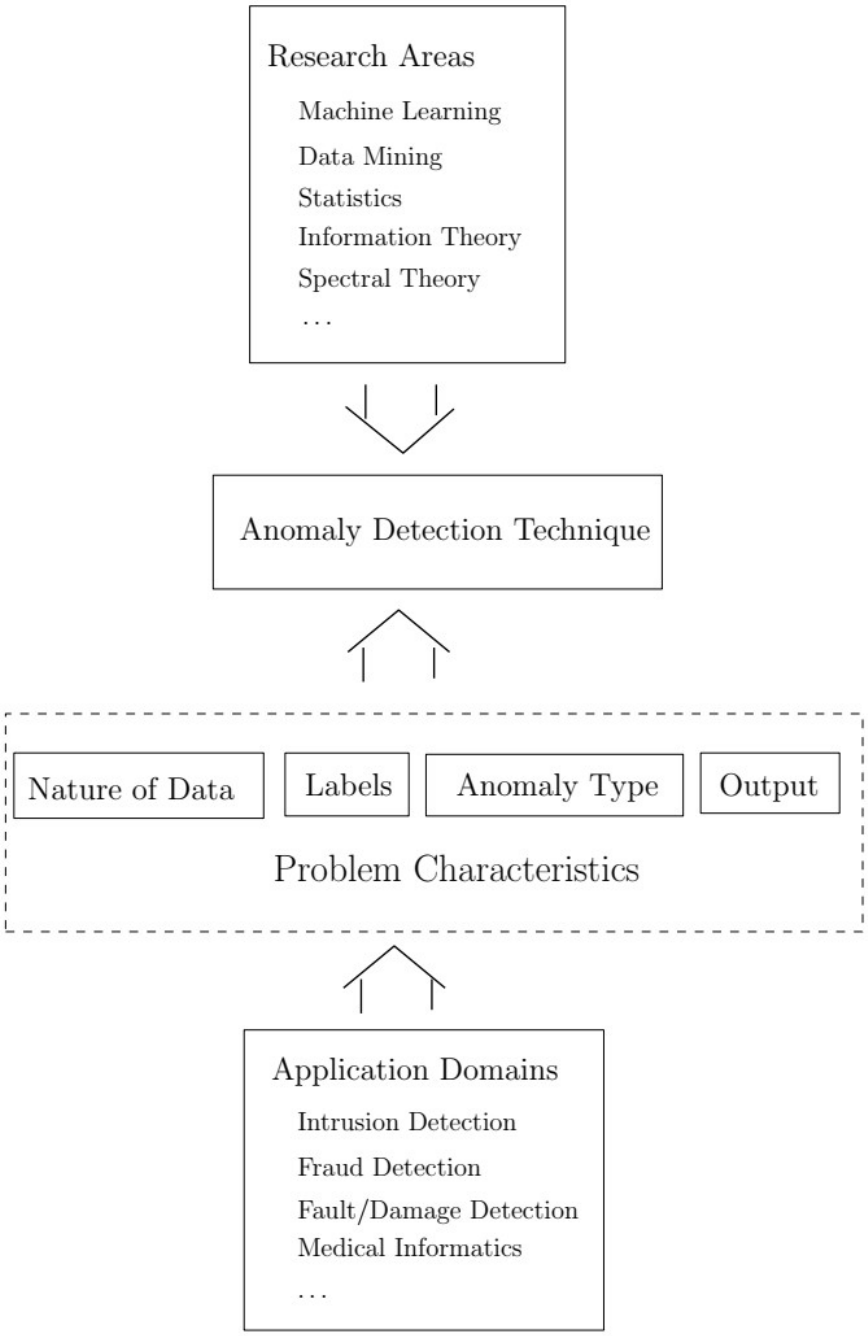
In 1977 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels

Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations? The researchers held back publishing their work for nearly a decade.

The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:
"Cosmic Imagery: Key Images in the History of Science"
By John D. Barrow
http://www3.epa.gov/ozone/science
https://ozonewatch.gsfc.nasa.gov/



Key components associated with an anomaly detection technique

Anomaly Detection : A Survey by
Varun Chandola, Arindam Banerjee
And Vipin Kumar University of Minnesota
ACM Computing Surveys, *September 2009*

More on Outlier/Anomaly Detection

Challenges

- How many outliers are there in the data?
- Method is unsupervised (sometimes supervised methods are used)
  - Validation can be quite challenging (just like for clustering)
- Finding needle in a haystack

Working assumption:

- There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

Outlier detection vs. *novelty detection* (identify new topics and trends in a timely manner in social media): early stage, outlier; but later merged into the model

Types of Outliers

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^{o} 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 (Contd.)

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

innovate achieve lead

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Outlier Detection Approaches

Outlier Detection 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

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

Mining Contextual Outliers: Transform into Conventional Outlier Detection

If the contexts can be clearly identified, transform it to conventional outlier detection

1. Identify the context of the object using the contextual attributes
2. Calculate the outlier score for the object in the context using a conventional outlier detection method

Ex. Detect outlier customers in the context of customer groups

- Contextual attributes: *age group*, *postal code*
- Behavioral attributes: *# of trans/yr*, *annual total trans. amount*

Steps:

(1) locate c's context,

(2) compare c with the other customers in the same group, and

(3) use a conventional outlier detection method

Mining Contextual Outliers: Modeling Normal Behavior with Respect to Contexts

In some applications, one cannot clearly partition the data into contexts

- Ex. if a customer suddenly purchased a product that is unrelated to those she recently browsed, it is unclear how many products browsed earlier should be considered as the context

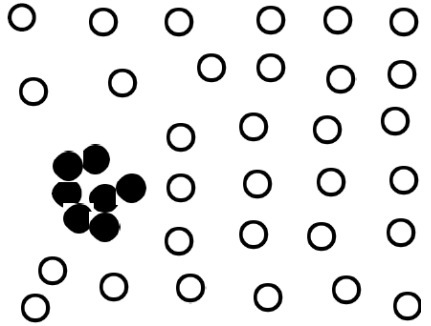
Model the "normal" behavior with respect to contexts

- Using a training data set, train a model that predicts the expected behavior attribute values with respect to the contextual attribute values
- An object is a contextual outlier if its behavior attribute values significantly deviate from the values predicted by the model

Using a prediction model that links the contexts and behavior, these methods avoid the explicit identification of specific contexts

Methods: A number of classification and prediction techniques can be used to build such models, such as regression, Markov Models, and Finite State Automaton

Mining Collective Outliers : On the Set of "Structured Objects"



- Collective outlier - objects as a group deviate from the entire data
- Need to examine the *structure* of the data set,
  - i.e, the relationships between multiple data objects
- Each of these structures is inherent to its respective type of data
  - For temporal data (such as time series and sequences)
    - explore the structures formed by time, which occur in segments of the time series or subsequences
  - For spatial data, explore local areas
  - For graph and network data, we explore subgraphs

Mining Collective Outliers : On the Set of "Structured Objects"

- Difference from the contextual outlier detection: the structures are often not explicitly defined, and have to be discovered as part of the outlier detection process.
- Collective outlier detection methods: two categories
  - Reduce the problem to conventional outlier detection
    - Identify *structure units*, treat each structure unit (e.g., subsequence, time series segment, local area, or subgraph) as a data object, and extract features
    - Then outlier detection on the set of "structured objects" constructed as such using the extracted features
    - e.g. Detect collective outliers in online social network of customers
      - Treat each possible subgraph of the network as a structure unit
      - Collective outlier: An outlier subgraph in the social network
        - Small subgraphs that are of very low frequency
        - Large subgraphs that are surprisingly frequent

Mining Collective Outliers II: Direct Modeling of the Expected Behavior of Structure Units

Model the expected behavior of structure units directly

- e.g. Detect collective outliers in temporal sequences
  - Learn a Markov model from the sequences
  - A subsequence can then be declared as a collective outlier if it significantly deviates from the model

Collective outlier detection is subtle due to the challenge of exploring the structures in data

- The exploration typically uses heuristics, and thus may be application dependent
- The computational cost is often high due to the sophisticated mining process

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

innovate achieve lead

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Statistical Outliers

Statistical Approaches

Statistical approaches assume that the objects in a data set are generated by a stochastic process (a generative model)

- The effectiveness of statistical methods highly depends on whether the assumptions made for the statistical model hold true for the given data

Statistic models used in the methods may be parametric or nonparametric.

- A **parametric method** assumes that the normal data objects are generated by a parametric distribution
- A **nonparametric method** does not assume an a priori statistical model. Instead, a nonparametric method tries to determine the model from the input data.

Discordancy test

The statistical distribution-based approach identifies outliers with respect to the model using a *discordancy test*.

A statistical discordancy test examines first a *working hypothesis*. A **working hypothesis**, *H*, is a statement that the entire data set of *n* objects comes from an initial distribution model, *F*, that is,

$H : o_i \in F$, where i=1,2,....n

The hypothesis is retained if there is no statistically significant evidence supporting its rejection

A **discordancy test** verifies whether an object, $\boldsymbol{o}_i$, is significantly large (or small) in relation to the distribution *F*

The result is very much dependent on which model *F* is chosen because $\boldsymbol{o}_i$ may be an outlier under one model and a perfectly valid value under another.

Alternative distributions

Inherent alternative distribution: In this case, the working hypothesis that all of the objects come from distribution *F* is rejected in favor of the alternative hypothesis that all of the objects arise from another distribution, *G*:

$$H_a : o_i \in G, \text{ where } i=1,2,....n$$

F and *G* may be different distributions or differ only in parameters of the same distribution. For example, it may have a different mean or dispersion, or a longer tail.

Alternative distributions

Mixture alternative distribution: The mixture alternative states that discordant values are not outliers in the *F* population, but contaminants from some other population, *G*. In this case, the alternative hypothesis is

$$H_a : o_i \in (1-\lambda)F + \lambda G \text{ , where } i=1,2,....n$$

Slippage alternative distribution: This alternative states that all of the objects (apart from some prescribed small number) arise independently from the initial model, *F*, with its given parameters, whereas the remaining objects are independent observations from a modified version of *F* in which the parameters have been shifted.

Statistical Approaches – Parametric Methods

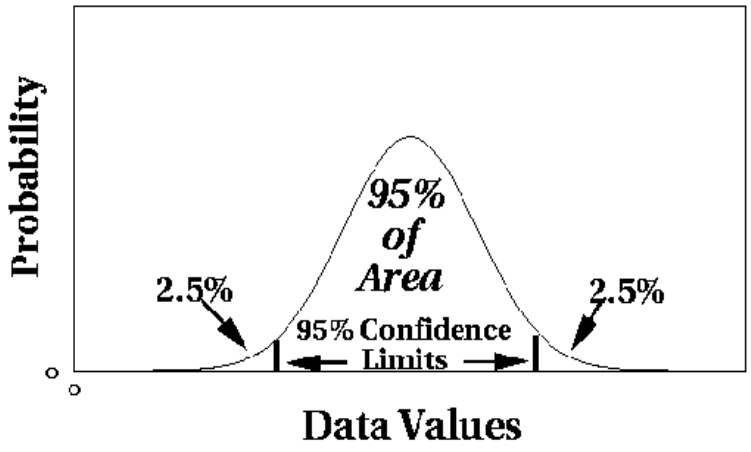
Assumes that the normal data is generated by a parametric distribution with parameter θ

The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution

The smaller this value, the more likely x is an outlier

The parametric distribution can be normal distribution with a mean and variance.

Outliers are points where probability of occurrence is below a threshold.



Parametric Methods: Univariate Outliers

- 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

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

Parametric Methods: Univariate Outliers

Avg. temp.: {24.0, 28.9, 28.9, 29.0, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4}

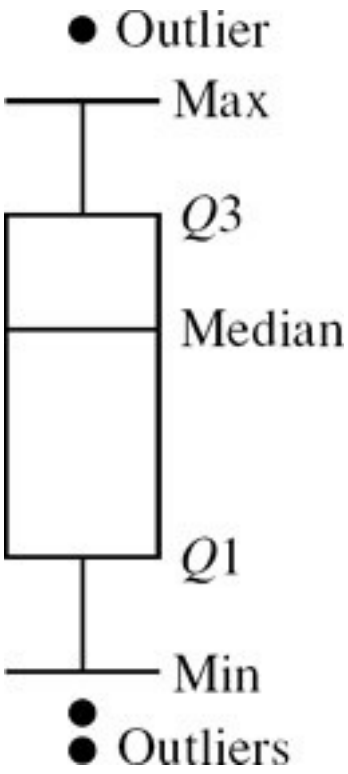
- For the above data with n = 10, we have

$$\hat{\mu} = 28.61 \;\; \hat{\sigma} = \sqrt{2.29} = 1.51$$

- Then (24 – 28.61) /1.51 = – 3.04 < –3, 24 is an outlier since

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

Visual Approach



A straightforward method for statistical outlier detection can also be used in visualization, e.g., the *boxplot method* plots the univariate input data using a five-number summary

the smallest nonoutlier value (Min),

the lower quartile (*Q*1),

the median (*Q*2),

the upper quartile (*Q*3), and

the largest nonoutlier value (Max).

The *interquantile range* (*IQR*) is defined as *Q*3 − *Q*1. Any object that is more than 1.5 × *IQR* smaller than *Q*1 or 1.5 × *IQR* larger than *Q*3 is treated as an outlier because the region between *Q*1 − 1.5 × *IQR* and *Q*3 + 1.5 × *IQR* contains 99.3% of the objects. The rationale is similar to using 3σ as the threshold for normal distribution

Parametric Methods: 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 Mahalanobis distance

- Mahalanobis distance is a measure of the distance between a point P and a distribution D.
- This distance is zero if P is at the mean of D, and grows as P moves away from the mean: along each principal component axis, it measures the number of standard deviations from P to the mean of D

Method 2. Use χ^2 –statistic:

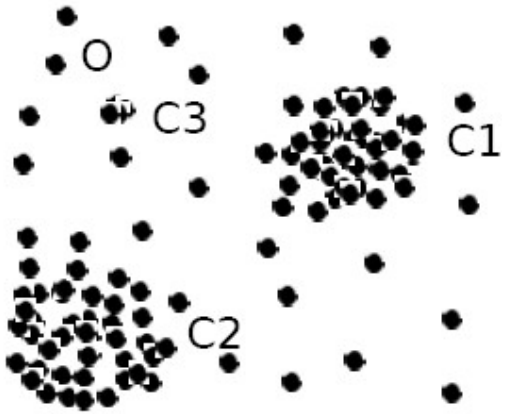
- where E_i is the mean of the i-dimension among all objects, and n is the dimensionality
- If χ^2 –statistic is large, then object o_i is an outlier

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$$

Parametric Methods: Using Mixture of Parametric Distributions

Assuming data generated by a normal distribution could be sometimes overly simplified

Example (right figure): The objects between the two clusters cannot be captured as outliers since they are close to the estimated mean



- To overcome this problem, assume the normal data is generated by two normal distributions. For any object o in the data set, the probability that o is generated by the mixture of the two distributions is given by

$$Pr(o|\Theta_1, \Theta_2) = f_{\Theta_1}(o) + f_{\Theta_2}(o)$$

  - where $f_{\theta 1}$ and $f_{\theta 2}$ are the probability density functions of θ_1 and θ_2
- Then use EM algorithm to learn the parameters μ_1, σ_1, μ_2, σ_2 from data
- An object o is an outlier if it does not belong to any cluster

Detecting outliers

There are two basic types of procedures for detecting outliers:

Block procedures: In this case, either all of the suspect objects are treated as outliers or all of them are accepted as consistent.

Consecutive (or sequential) procedures: e.g. *inside-out* procedure. The idea is that the object that is least "likely" to be an outlier is tested first. If it is found to be an outlier, then all of the more extreme values are also considered outliers; otherwise, the next most extreme object is tested, and so on.

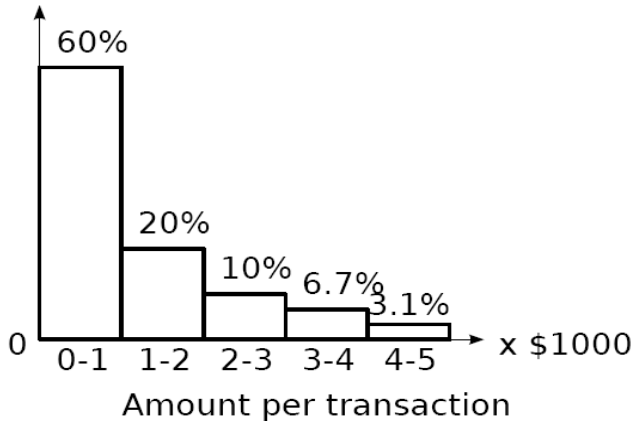
(This procedure tends to be more effective than block procedures.)

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

innovate achieve lead

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Proximity Based Outliers

Proximity-Based Approaches: Distance-Based vs. Density-Based Outlier Detection

Intuition: Objects that are far away from the others are outliers

Assumption of proximity-based approach: The proximity of an outlier deviates significantly from that of most of the others in the data set

Two types of proximity-based outlier detection methods

- Distance-based outlier detection: An object o is an outlier if its neighborhood does not have enough other points
- Density-based outlier detection: An object o is an outlier if its density is relatively much lower than that of its neighbors

Distance-Based Outlier Detection

For each object o, examine the # of other objects in the *r*-neighborhood of o, where *r* is a user-specified **distance threshold**

An object o is an outlier if most (taking π as a **fraction threshold**) of the objects in D are far away from o, i.e., not in the r-neighborhood of o

An object o is a DB(r, π) outlier if

$$\frac{\|\{o' | dist(o, o') \leq r\}\|}{\|D\|} \leq \pi$$

Equivalently, one can check the distance between *o* and its *k*-th nearest neighbor o_k, where

$$k = \lceil \pi \|D\| \rceil$$

o is an outlier if dist(o, o_k) > r

Distance-Based Outlier Detection

Algorithm: Distance-based outlier detection.

Input:

- a set of objects $D = \{o_1, \ldots, o_n\}$, threshold r $(r > 0)$ and π $(0 < \pi \leq 1)$;

Output: $DB(r, \pi)$ outliers in D.

Method:

```
for i = 1 to n do
    count ← 0
    for j = 1 to n do
        if i ≠ j and dist(o_i, o_j) ≤ r then
            count ← count + 1
            if count ≥ π · n then
                exit {o_i cannot be a DB(r, π) outlier}
            endif
        endif
    endfor
    print o_i {o_i is a DB(r, π) outlier according to (Eq. 12.10)}
endfor;
```

Efficient computation: Nested loop algorithm

- For any object o_i, calculate its distance from other objects, and count the # of other objects in the r-neighborhood.
- If $\pi \cdot n$ other objects are within r distance, terminate the inner loop
- Otherwise, o_i is a DB(r, π) outlier

Efficiency: Actually CPU time is not $O(n^2)$ but linear to the data set size since for most non-outlier objects, the inner loop terminates early

Distance-Based Outlier Detection: Improving Algorithm

Why efficiency is still a concern? When the complete set of objects cannot be held into main memory, cost of I/O swapping will be high

The major cost:

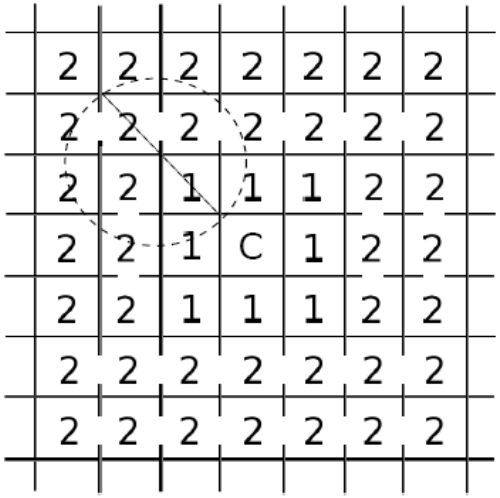
(1) each object tests against the whole data set, why not only its close neighbor?

(2) instead of checking objects one by one, why not group by group?

Grid-based method (CELL): Data space is partitioned into a multi-D grid. Only adjoining cells are checked for determining if object is an outlier

Distance-Based Outlier Detection: A Grid-Based Method

Grid-based method (CELL): Data space is partitioned into a multi-D grid. Each cell is a hyper cube with diagonal length r/2



- Pruning using the level-1 & level 2 cell properties:
  - For any possible point x in cell C and any possible point y in a level-1 cell, dist(x,y) ≤ r
  - For any possible point x in cell C and any point y such that dist(x,y) ≥ r, y is in a level-2 cell

- Thus we only need to check the objects that cannot be pruned, and even for such an object o, only need to compute the distance between o and the objects in the level-2 cells (since beyond level-2, the distance from o is more than r)

Distance-Based Outlier Detection: Limitations

Distance-based outliers, such as $DB(r,\pi)$-outliers, are just one type of outlier

Distance-based outlier detection takes a global view of the data set

$DB(r,\pi)$-outlier, for example, is far (as quantified by parameter r) from at least $(1-\pi)\times 100\%$ of the objects in the data set. In other words, an outlier as such is remote from the majority of the data.

To detect distance-based outliers, we need two global parameters, *r* and π, which are applied to every outlier object.

Many real-world data sets demonstrate a more complex structure, where objects may be considered outliers with respect to their local neighborhoods, rather than with respect to the global data distribution.

innovate achieve lead

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

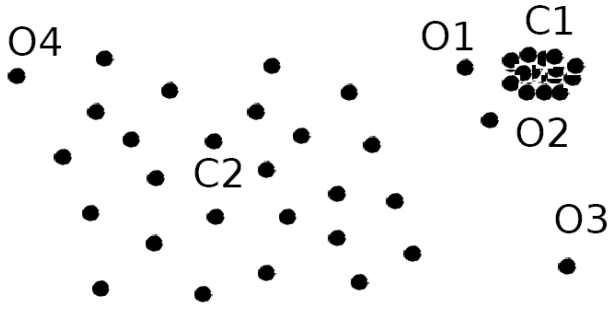
Density-Based Outlier

Density-Based Outlier Detection

Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution

In Fig., o_1 and o2 are local outliers to C_1, o_3 is a global outlier, but o_4 is not an outlier. However, proximity-based clustering cannot find o_1 and o_2 are outlier (e.g., comparing with O_4).

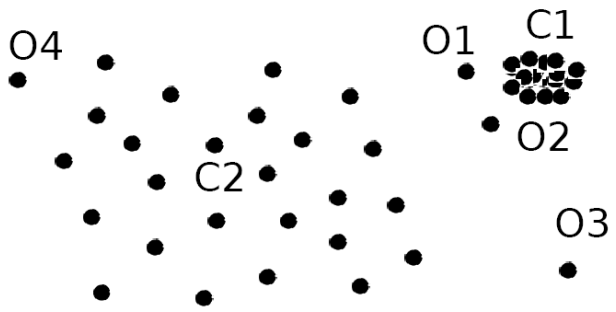
- Intuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbors
- Method: Use the relative density of an object against its neighbors as the indicator of the degree of the object being outliers



Density-Based Outlier Detection

Local outliers: Outliers comparing to their local neighborhoods, instead of the global data distribution

In Fig., o_1 and o2 are local outliers to C_1, o_3 is a global outlier, but o_4 is not an outlier. However, proximity-based clustering cannot find o_1 and o_2 are outlier (e.g., comparing with O_4).



- Intuition (density-based outlier detection): The density around an outlier object is significantly different from the density around its neighbors
- Method: Use the relative density of an object against its neighbors as the indicator of the degree of the object being outliers

Local Reachability Density

k-distance of an object o, $dist_k(o)$: distance between o and its k-th NN

k-distance neighborhood of o, $N_k(o) = \{o' \mid o' \text{ in } D, dist(o, o') \leq dist_k(o)\}$

- $N_k(o)$ could be bigger than k since multiple objects may have identical distance to o

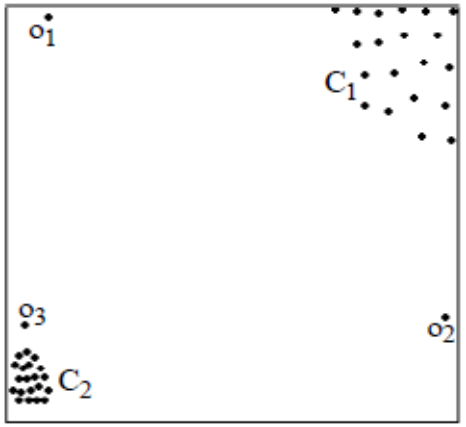
Reachability distance from *o'* to *o*:

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$$

where k is a user-specified parameter
that controls the smoothing effect

Local reachability density of *o*:

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$



Local Outlier Factor: LOF

- LOF (Local outlier factor) of an object o is the average of the ratio of local reachability of *o* and those of *o*'s k-nearest neighbors

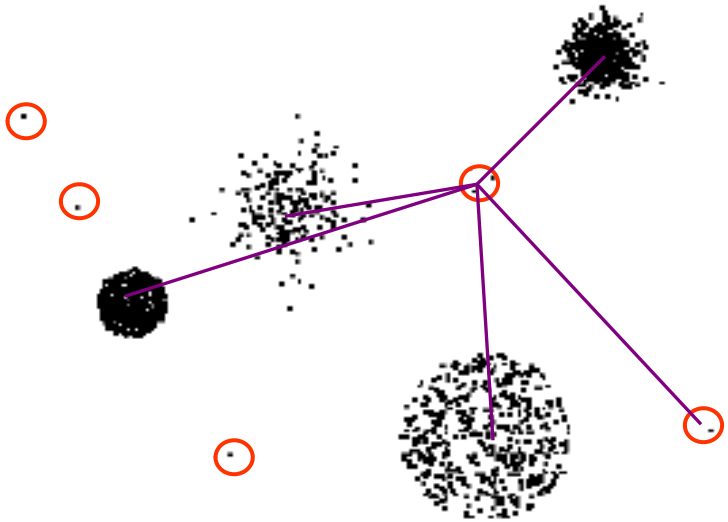
$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|}$$

- the local outlier factor is the average of the ratio of the local reachability density of o and those of o's k-nearest neighbors
- The lower the local reachability density of o, and the higher the local reachability density of the k-NN of o, the higher LOF
- This captures a local outlier whose local density is relatively low comparing to the local densities of its k-NN

Clustering-Based

Basic idea:

- Cluster the data into groups of different density
- Choose points in small cluster as candidate outliers
- Compute the distance between candidate points and non-candidate clusters.
  - If candidate points are far from all other non-candidate points, they are outliers



innovate achieve lead

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Base Rate Fallacy – An outlier challenge

Base Rate Fallacy

- Bayes theorem:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

- More generally:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}$$

Base Rate Fallacy (Axelsson, 1999)

The base-rate fallacy is best described through example. Suppose that your doctor performs a test that is 99% accurate, i.e. when the test was administered to a test population all of whom had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative. Upon visiting your doctor to learn the results he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that out of the entire population the rate of incidence is only 1/10000, i.e. only 1 in 10000 people have this ailment. What, given this information, is the probability of you having the disease?

The Base-Rate Fallacy and the Difficulty of Intrusion Detection by Stefan Axelsson ACM Transactions on Information and System Security

Base Rate Fallacy

$$P(S|P) = \frac{P(S) \cdot P(P|S)}{P(S) \cdot P(P|S) + P(\neg S) \cdot P(P|\neg S)}$$

$$P(S|P) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = \\ = 0.00980\ldots \approx 1\%$$

- Even though the test is 99% certain, your chance of having the disease is 1/100, because the population of healthy people is much larger than sick people

Base Rate Fallacy in Intrusion Detection

I: intrusive behavior,

▲I: non-intrusive behavior

A: alarm

▲A: no alarm

Detection rate (true positive rate): P(A|I)

False alarm rate: P(A|▲I)

Goal is to maximize both

- Bayesian detection rate, P(I|A)
- P(▲I|▲A)

Detection Rate vs False Alarm Rate

$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)}$$

Suppose there are about 20 intrusions Per million

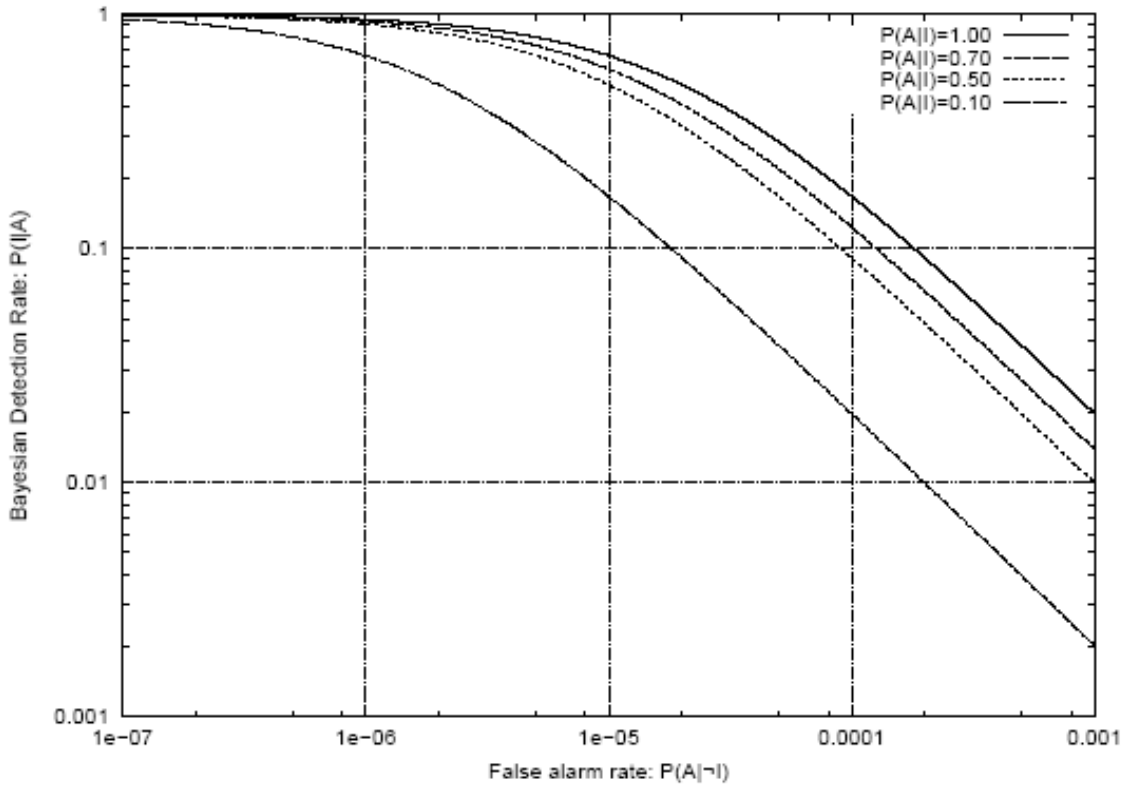
$$P(I) = 1 \Big/ \frac{1 \cdot 10^6}{2 \cdot 10} = 2 \cdot 10^{-5};$$

$$P(\neg I) = 1 - P(I) = 0.99998$$

$$P(I|A) = \frac{2 \cdot 10^{-5} \cdot P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.99998 \cdot P(A|\neg I)}$$

- False alarm rate becomes more dominant if P(I) is very low

Detection Rate vs False Alarm Rate



- Axelsson: We need an extremely low false alarm rate to achieve a reasonable Bayesian detection rate

Prescribed Text Books

	Author(s), Title, Edition, Publishing House
T1	Tan P. N., Steinbach M & Kumar V. “Introduction to Data Mining” Pearson Education
T2	Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber and Jian Pei Morgan Kaufmann Publishers
R1	Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner by Vijay Kotu and Bala Deshpande Morgan Kaufmann Publishers
	The Base-Rate Fallacy and the Difficulty of Intrusion Detection by Stefan Axelsson ACM Transactions on Information and System Security, Vol. 3, No. 3, August 2000, Pages 186–205