

CSI4142 Introduction to Data Science

Detecting Outliers

(Slides by HL Viktor, based on Chapters 12 of Han et. al.)

1

Detecting Outliers

- What are Outliers?
- What are the types of Outliers?
- Outlier Detection Methods
 1. Statistical Approaches
 2. Proximity-Base Approaches
 3. Clustering-Base Approaches
 4. Classification Approaches
- Mining Contextual and Collective Outliers
- Outlier Detection in High Dimensional Data
- Summary



2

What Are Outliers?

- **Outlier:** A data object that **deviates significantly from the normal objects** as if it were **generated by a different mechanism**
 - Ex.: Unusual credit card purchase, sports: Michael Jordon, Wayne Gretzky, ...
- **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



3

What are outliers?

- Outlier detection vs. *novelty detection*: early stage, outlier; but **later merged into the model**
- **Applications:**
 - Credit card fraud detection
 - Telecom fraud detection
 - Customer segmentation
 - Medical analysis



4

Types of Outliers

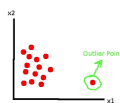
Three kinds:

global, contextual and collective outliers

5

Global Outliers

- Also known as **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 (what is abnormal?)



6

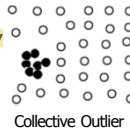
Contextual outliers



- Also known as **conditional outlier**
 - Object is O_c if it **deviates significantly based on a selected context**
 - Temperature of 25C in Ottawa an outlier???
 - 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?

7

Collective Outliers



Collective Outlier

- 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

8

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**

9

Outlier Detection Methods

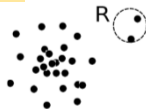
1. Statistical Approaches
2. Proximity-Base Approaches
3. Clustering-Base Approaches
4. Classification Approaches



10

1. Statistical Methods

- Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model (a stochastic model)
 - The data not following the model are outliers.
- Example (right figure): First use Gaussian distribution to model the normal data
 - For each object y in region R , estimate $g_0(y)$, the probability of y fits the Gaussian distribution
 - If $g_0(y)$ is very low, y is unlikely generated by the Gaussian model, thus an outlier
- 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
 - E.g., parametric vs. non-parametric



11

a. Parametric Methods: Grubb's Test

- **Univariate outlier detection: Grubb's test** (maximum normed residual test)
 - a statistical method under normal distribution
 - For each object x in a data set, compute its z-score: x is an outlier if

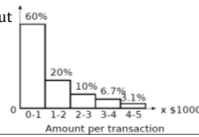
$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t_{\alpha/(2N), N-2}^2}{N-2 + t_{\alpha/(2N), N-2}^2}}$$

where $t_{\alpha/(2N), N-2}^2$ is the value taken by a t-distribution at a significance level of $\alpha/(2N)$, and N is the # of objects in the data set

12

b. 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 \rightarrow normal objects in empty/rare bins, false positive
 - Too big bin size \rightarrow outliers in some frequent bins, false negative
 - Solution:** Adopt kernel density estimation to estimate the probability density distribution of the data. If the estimated density function is high, the object is likely normal. Otherwise, it is likely an outlier.



13

2. Proximity-Based Approaches

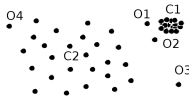
- Intuition: Objects that are far away from the others are outliers
- Assumption of proximity-based approach:
 - The proximity of an outlier deviates significantly from that of most of the others in the data set
- Two types of proximity-based outlier detection methods
 - Distance-based outlier detection:** 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



14

Density-Based Outlier Detection

- Local outliers:** Outliers comparing to their local neighborhoods, instead of the global data distribution
- In Fig., o_1 and o_2 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
- k-distance** of an object o , $\text{dist}_k(o)$: distance between o and its k -th NN
- k-distance neighborhood** of o , $N_k(o) = \{o' \mid o' \text{ in } D, \text{dist}(o, o') \leq \text{dist}_k(o)\}$
 - $N_k(o)$ could be bigger than k since multiple objects may have identical distance to o



15

3. Clustering-Based Methods

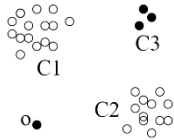
- Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters
- Example (right figure): two clusters
 - All points not in R form a large cluster
 - The two points in R form a tiny cluster, thus are outliers
- Since there are many clustering methods, there are many clustering-based outlier detection methods as well



16

Clustering-Based Outlier Detection (3): Detecting Outliers in Small Clusters

- **FindCBLOF**: Detect outliers in small clusters
 - Find clusters, and sort them in decreasing size
 - To each data point, assign a **cluster-based local outlier factor (CBLOF)**:
 - If obj p belongs to a large cluster, $CBLOF = \text{cluster_size} \times \text{similarity between } p \text{ and cluster}$
 - If p belongs to a small one, $CBLOF = \text{cluster size} \times \text{similarity between } p \text{ and the closest large cluster}$



- Ex. In the figure, o is outlier since its closest large cluster is C_1 , but the similarity between o and C_1 is small. For any point in C_3 , its closest large cluster is C_2 but its similarity from C_2 is low, plus $|C_3| = 3$ is small

17

17

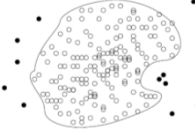
4. Classification-based methods

- One class learning
- Semi-supervised learning




18

Method I: One-Class Model

- Idea: Train a classification model that can distinguish "normal" data from outliers
 - A brute-force approach: Consider a training set that contains samples labeled as "normal" and others labeled as "outlier"
 - But, the training set is typically heavily biased: # of "normal" samples likely far exceeds # of outlier samples
 - Cannot detect unseen anomaly
- 
- One-class model: A classifier is built to describe only the normal class.
 - Learn the decision boundary of the normal class using classification methods such as SVM
 - Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers
 - Neat idea: Can detect new outliers that may not appear close to any outlier objects in the training set
 - Extension: Normal objects may belong to multiple classes

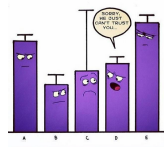
19

Method II: Semi-Supervised Learning

- Semi-supervised learning: Combining classification-based and clustering-based methods
 - Method
 - Using a clustering-based approach, find a large cluster, C , and a small cluster, C_1
 - Since some objects in C carry the label "normal", treat all objects in C as normal
 - Use the one-class model of this cluster to identify normal objects in outlier detection
 - Since some objects in cluster C_1 carry the label "outlier", declare all objects in C_1 as outliers
 - Any object that does not fall into the model for C (such as a) is considered an outlier as well
- 
- Comments on classification-based outlier detection methods
 - Strength: Outlier detection is fast
 - Bottleneck: Quality heavily depends on the availability and quality of the training set, but often difficult to obtain representative and high-quality training data

20

Mining Contextual and Collective Outliers



21

Mining Contextual Outliers

- 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

22

Mining Collective Outliers

- Models the expected behavior of structure units directly
- Ex. 1. 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
- Ex. 2. 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

23

Finding Outliers in High Dimensions



24

Challenges for Outlier Detection in High-Dimensional Data

- Interpretation of outliers
 - Detecting outliers without saying why they are outliers is not very useful in high-D due to many features (or dimensions) are involved in a high-dimensional data set
 - E.g., which subspaces that manifest the outliers or an assessment regarding the “outlier-ness” of the objects
- Data sparsity
 - Data in high-D spaces are often sparse
 - The distance between objects becomes heavily dominated by noise as the dimensionality increases
- Data subspaces
 - Adaptive to the subspaces signifying the outliers
 - Capturing the local behavior of data
- Scalable with respect to dimensionality
 - # of subspaces increases exponentially

25

Approach: Finding Outliers in Subspaces

- Find outliers in much lower dimensional subspaces: easy to interpret why and to what extent the object is an outlier
 - E.g., find outlier customers in certain subspace: *average transaction amount >> avg. and purchase frequency << avg.*
- Ex. A grid-based subspace outlier detection method
 - Project data onto various subspaces to find an area whose density is much lower than average
 - Discretize the data into a grid
 - Search for regions that are significantly sparse

26

Summary

- Types of outliers
 - global, contextual & collective outliers
- Outlier detection
 1. Statistical (or model-based) approaches
 2. Proximity-base approaches
 3. Clustering-base approaches
 4. Classification approaches
- Mining contextual and collective outliers
- Outlier detection in high dimensional data



27

We are done!

28
