4.1 Outlier Detection

Basic Concepts

- **Outlier**: represent patterns in data that do not conform to a well-defined notion of normal
- They can represent:
    - Errors
    - Truthful deviation of data
    - Critical information that can trigger preventive or corrective information

Outliers and Anomalies

- Roughly related
- **Outliers** can have a negative connotation, being associated with data noise
- **Anomalies** are often associated with unusual data that should be further investigated to identify the cause of occurrence
- Anomaly can be considered as an outlier, but an outlier is not necessarily an anomaly
- The following outlier detection application and methods involve outliers that can be seen as anomalies

Application of Outlier Detection

- **Quality Control and Fault Detection Applications**: Quality Control, Fault Detection and Systems Diagnosis, Structure Defect Detection
- **Financial Applications**: Credit Card Fraud, Insurance Claim Fraud, Stock Market Anomalies, Financial Interaction Networks
- **Intrusion and Security Applications**: Host-based Intrusions, Network Intrusion Detection
- **Web Log Analytics**: Web Log Anomalies
- **Market Basket Analysis**: Outlier transactions in association patterns
- **Medical Applications**: Medical Sensor Diagnostics, Medical Imaging Diagnostics
- **Text and Social Media Applications**: Event Detection in Text and Social Media, Spam Email, Noisy and Spam Links, Anomalous Activity in Social Networks
- **Earth Science Applications**: Sear Surface Temperature Anomalies, Land Cover Anomalies, Harmful Algae Blooms

Challenges of Outlier Detection

- Define every possible "normal" behavior is hard
- The boundary between normal and an outlying behavior is often not precise

- There is no general outlier definition; it depends on the application domain
- It is difficult to distinguish real meaningful outliers from simple random noise in data
- The outlier behavior may evolve with time
- Malicious actions adapt themselves to appear as normal
- Inherent lack of known labeled outliers for training/validation of models

Key Aspects of Outlier Detection Problem

- Nature of Input Data
- Type of Outliers
- Intended Output
- Learning Task
- Performance Metrics

Nature of Input Data

- **Univariate**: one attribute
- **Multivariate**: multiple attributes

Relationship among data instances:

- None

- Sequential/Temporal

- Spatial

- Spatio-temporal

- Graph

- Dimensionality of data

Types of Outliers

Point Outlier

An instance that individually or in small groups is very different from the rest of the instances

Contextural Outliers

An instance that when considered within a context is very different from the rest of the instances

Collective Outlier

An instance that, even though individually may not be an outlier, inspected in conjunction with related instances and with respect to the entire data set is an outlier

Intended Output

Assign a:

- **label/value**: identification normal or outlier instance
- **score**: probability of being an outlier (allow the output to be ranked; required the specification of a threshold)

Learning Task

Unsupervised Outlier Detection

- Data set has no information on the behavior of each instance
- It assumes that instances with normal behavior are far more frequent
- Most common case in real-life applications

Semi-supervised Outlier Detection

- Data set has a few instances of normal or outlier behavior
- Some real-life applications, such as fault detection, provide such data

Supervised Outlier Detection

- Data set has instances of both normal and outlier behavior
- Hard to obtain such data in real-life applications

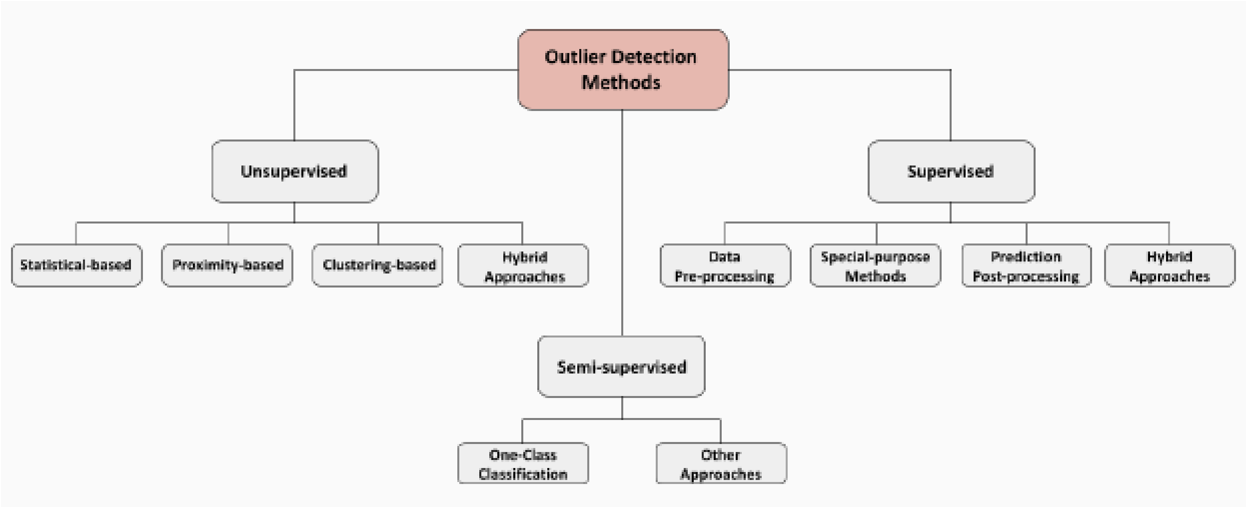
Performance Metrics

Inadequacy of Standard Performance Metrics

- They assume that all instances are equally relevant for the model performance
- These metrics would give a good performance estimation to a model that performs well on normal (frequent) cases and bad on outlier (rare) cases

Outlier Detection Approaches

Taxonomy of Outlier Detection Methods



Unsupervised Learning Techniques

Statistical-based

- **Proposal**: all the points that satisfy a statistical discordance test for some statistical model are declared as outliers
- **Advantages**:
  - If the assumptions of the statistical model hold true, these techniques provide a justifiable solution for outlier detection
  - The outlier score is associated with a confidence interval
- **Techniques**: Parametric, Non-parametric

Parametric Techniques

Assume one of the known probability distribution functions:

- **Grubbs' Test**: a statistical test used to detect outliers in a **univariate** data set assumed to come from a normally distributed population
- **Boxplot**: assumes a near-normal distribution of the values in a **univariate** data set, and identifies as outlier any value outside the interval
- **Mahalanobis distance**:
  - assumes a **multivariate** normal distribution of data
  - incorporates dependencies between attributes by the covariance matrix
  - transforms a multivariate outlier detection task into a univariate outlier detection problem
  - large Mahalanobis distance = outlier
- **Mixture of parametric distributions**

Non-parametric Techniques

Probability distribution function is not assumed, but estimated from data

- **Histograms**: used for both univariate and multivariate data. For the latter, the attribute-wise histograms are constructed, and an aggregated score is obtained.

Hard to choose the appropriate bin size
- **Kernel functions**: adopt a kernel density estimation to estimate the probability density distribution of the data. Outliers are in regions with low density

Disadvantages

- The data does not always follow a statistical model
- Choosing the best hypothesis test statistics is not straightforward
- Capture interactions between attributes is not always possible
- Estimating the parameters for some statistical models is hard

Proximity-based

- **Proposal**: normal instances occur in dense neighborhoods, while outliers occur far from their closest neighbors
- **Advantages**:
  - Purely data driven technique
  - Does not make any assumptions regarding the underlying distribution of data
- **Techniques**: Distance-based, Density-based

Distance-based

Case c is an outlier if less than k cases are within a distance λ of c

- Define proper distance metric
- Define a "reasonable" neighborhood (λ)
- Define what is a "lot of other points" (k)
- **Major cost**: for each point is calculated its distance to all the other points
- **Use of Global Distance**: measures poses difficulties in detecting outliers in data sets with different density regions

Density-based

- Outliers should be **locally** inspected
- Compare points to their local neighborhood, instead of the global data distribution
- The density around an outlier is significantly different from the density around its neighbors
- Use the **relative density** of a point against its neighbors as the indicator of the degree of the point being an outlier
- Outliers are points in lower local density areas with respect to the density of its local neighborhood
- **LOF (Local Outlier Factor)**:
  - **k-distance**: distance between p and its k-th nearest neighbor

- **k-distance neighborhood**: all the points whose distance from p is not greater than the k-distance
- **reachability-distance of p with respect to o**: the maximum between their k-distance and their actual distance
- **intuition**: high values of reachability-distance between 2 give points indicated that they may not be in the same cluster
- **local reachability-density**: inversely proportional to the average reachability-distance of its k neighborhood
- lower local reachability-density in comparison to its k-neighborhood → high value

Disadvantages

- True outliers and noisy regions of low density may be hard to distinguish
- These methods need to combine global and local analysis
- In high dimensional data, the contrast in the distances is lost
- Computational complexity of the test phase

Clustering-based

- **Proposal**: normal instances belong to large and dense clusters, while outlier instances are instances that
  - do not belong to any of the clusters
  - are far from its closest cluster
  - form very small or low density clusters
- **Advantages**:
  - easily adaptable to online/incremental mode
  - test phase is fast
- **Techniques**: DBSCAN, FindCBLOF, ORH

DBSCAN

- Clustering method based on the notion of "density" of the points
- The density of a point is estimated by the number of points that are within a certain radius
- Based on this idea, points can be classified as:
  - **core points**: it the number of points within its radius are above a threshold
  - **border points**: if the number of points within its radius are not above a threshold, but they are within a radius of a core point
  - **noise points**: if do not have enough points within their radius, nor are sufficiently close to any core point
- noise points are removed for the formation of clusters
- all core points that are within a certain distance of each other are allocated to the same cluster
- each border point is allocated to the cluster of the nearest core points

- noise points are identified as outliers

FindCBLOF

- to each point, assign a cluster-based local outlier factor (CBLOF)
- the CBLOF score of a point p is determined by the size of the cluster to which p belongs and the distance between p and
    - its cluster centroid, if p belong to a large cluster
    - its closest latge cluster centroid, if p belongs to a small cluster

ORH

- obtain an agglomerative hierarchical clustering of the data set
- use the information on the "path" of each point through the dendogram as a form to determine its degree of outlyingness

Disadvantages

- Computationally expensive in the training phase
- If normal points do not create any clusters, this technique may fail
- In high dimensional spaces, clustering algorithms may not give any meaningful clusters
- Some techniques detect outliers as a byproduct

Isolation Forest

iForest detects outliers purely based on the concept of isolation without employing any distance or density measure.

- **Isolation**: separating an instance from the rest of the instances
- 2-stage process:
    1. The first (training) stage builds an ensemble of data-induced random binary decision trees (isolation trees) using sub-samples of the given training set.
    2. The second (evaluation) stage passes test instances through isolation trees to obtain an outlier score for each instance.
- **Parameters**: number of trees and subsampling size
- The score is related to average path length
    - outliers are more likely to be isolated closer to the root
    - normal points are more likely to be isolated at the deeper levels

Advantages

- No distance or density measures to detect anomalies
- Eliminates a major computational cost of distance calculation in all distance-based and density-based methods

- Scales up to handle extremely large data size and high-dimensional problems with many irrelevant attributes.

Disadvantages

- Hyperparameters that must be tuned
- Randomness component: different runs can give different results
- Large sample sizes may cause masking or swamping.

Semi-supervised Learning Techniques

One Class Classification

- **Proposal**: build a prediction model to the normal behavior and classify any deviations from this behavior as outliers
- **Advantages**:
  - Models are interpretable
  - Normal behaviour can be accurately learned
  - Can detect new outliers that may not appear close to any outlier points in the training set.
- **Techniques**: Auto-associative neural networks, One-class SVM

Auto-associative neural networks

- A feed-forward perceptron-based network is trained with normal data only
- The network has the same number of input and output nodes and a decreased number of hidden nodes to induce a bottleneck
- This bottleneck reduces the redundancies and focus on the key attributes of data
- After training, the output nodes recreate the example given as input nodes
- The network will successfully recreate normal data, but will generate a high-recreation error for outlier data.

One-class SVM

- It obtains a spherical boundary, in the feature space, around the normal data. The volume of this hypersphere is minimized, to minimize the effect of incorporating outliers in the solution
- The resulting hypersphere is characterized by a center and a radius R
- The optimization problem consists of minimizing the volume of the hypersphere, so that includes all the training points
- Every point lying outside this hypersphere is an outlier

Disadvantages

- Requires previous labeled instances for normal behaviour.

- Possible high false alarm rate - previously unseen normal data may be identified as an outlier.

Advanced Topics

Contextual Outlier Detection

- **Proposal**:
    - If a data instance is an outlier in a specific context (but not otherwise), then it is considered as a contextual outlier.
    - Each data instance is defined using two sets of attributes:
        - **Contextual attributes**: used to determine the context (or neighborhood) for that instance.
            - **Sequential Context**: position, time
            - **Spatial Context**: latitude, longitude
            - **Graph Context**: weights, edges.
        - **Behavioral attributes**: which define the non-contextual characteristics of an instance.
    - The outlier behaviour is determined using the values for the behavioural attributes within a specific context.
- **Advantages**:
    - Allow a natural definition of outlier in many real-life applications
    - Detects outliers that are hard to detect when analyzed from the global perspective.
- **Techniques**:
    - **Reduction to point outlier detection**
        - Segment data using contextual attributes
        - Apply a traditional point outlier within each context using behavioural attributes
        - Model “normal” behaviour with respect to contexts: an object is an outlier if its behavioural attributes significantly deviate from the values predicted by the model
    - **Utilizing structure in data**
        - Build models from the data using contextual attributes to predict the expected behaviour with respect to a given context
        - Avoids explicit identification of specific contexts
- **Disadvantages**
    - Identifying a set of good contextual attributes
    - It assumes that all normal instances within a context will be similar (in terms of behavioural attributes), while the outliers will be different.

Collective Outlier Detection

- **Proposal**:

    - If a collection of related data instances is anomalous with respect to the entire data set, then it is considered a collective outlier.
    - The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous.
- **Advantages**
    - Allow a natural definition of outlier in many real-life applications in which data instances are related.
- **Techniques**
    - A collective outlier can also be a contextual outlier if analyzed with respect to a context.
    - A collective outlier detection problem can be transformed to a contextual outlier detection problem by incorporating the context information.
- **Disadvantages**:
    - Contrary to contextual outliers, the structures are often not explicitly defined, and have to be discovered as part of the outlier detection process.
    - Need to extract features by examining the structure of the dataset, i.e. the relationship among data instances for:
        - sequence data to detect anomalous sequences
        - spatial data to detect anomalous sub-regions
        - graph data to detect anomalous sub-graphs.
    - The exploration of structures in data typically uses heuristics, and thus may be application dependent.
    - The computational cost is often high due to the sophisticated mining process.

Outlier Detection in High Dimensional Data

Challenges

- **Interpretation of outliers**
    - Detecting outliers without saying why they are outliers is not very useful in high-D due to the many features (or dimensions) involved
    - Identify the subspaces that manifest the outliers
- Data sparsity
    - Data in high-D spaces is often sparse
    - The distance between objects becomes heavily dominated by noise as the dimensionality increases
- Data subspaces
    - Capturing the local behavior of data
- Scalable with respect to dimensionality
    - nr. of subspaces increases exponentially

Techniques

- Find distance-based outliers, but use the ranks of distance instead of the absolute distance in outlier detection
- Dimensionality reduction: the principal components with low variance are preferred because, on such dimensions, normal objects are likely close to each other and outliers often deviate from the majority
- Project data onto various subspaces to find an area whose density is much lower than average.
- Develop new models for high-dimensional outliers directly. Avoid proximity measures and adopt new heuristics that do not deteriorate in high-dimensional data.

< Go back