Anomaly Detection - Basics

Huiping Cao

Outline

- General concepts
  - What are outliers
  - Types of outliers
  - Causes of anomalies
- Challenges of outlier detection
- Outlier detection approaches

What are outliers

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Assumption
  - There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

Anomaly/Outlier Detection

- Natural implication is that anomalies are relatively rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July
- Can be important or a nuisance
  - 10 foot tall 2 year old
  - Unusually high blood pressure

Applications

- Fraud detection (credit card usage)
- Intrusion detection (computer systems, computer networks)
- Ecosystem disturbances
- Public health
- Medicine

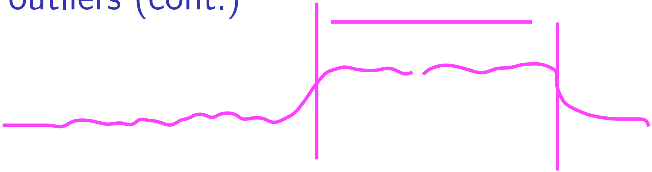
Types of outliers

- Global: deviate significantly from the rest of the dataset
  - Also called point anomalies
  - Most outlier detection methods are designed to find such outliers
- Example
  - Intrusion detection in network traffic

Types of outliers (cont.)

- Contextual (conditional) outliers
  - An object is an outlier in one context, but may be normal in another context
  - Contextual attributes: define the objects context
    - date, location
  - Behavior attributes: define the objects characteristics, and are used to evaluate whether the object is an outlier in the context.
    - temperature
  - A generalization of local outlier, defined in density based analysis
  - Background information to determine contextual attributes, etc.

Types of outliers (cont.)



- Collective: a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set
  - The individual data objects may not be outliers
  - Applications: supply-chain, web visiting, network (denial-of-service)
  - Need background information to make object relationships

Causes of Anomalies

- Data from different classes
  - Hawkins' definition of an outlier: an outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.
  - Measuring the weights of oranges, but a few grapefruit are mixed in
- Natural variation
  - Anomalies that represent extreme or unlikely variations
  - E.g., unusually tall people
- Data measurement and collection errors
  - Removing such anomalies is the focus of data preprocessing (data cleaning)
  - E.g., 200 pound 2 year old

Challenges of outlier detection

- Model normal/outlier objects
  - Hard to model complete normal behavior
  - Some methods assign "normal" or "abnormal"
  - Some methods assign a score measuring the "outlier-ness" of the object.
- Universal outlier detection: hard to develop
  - Similarity and distance definition is application-dependent
- Common issues: noise
- Understandability
  - Understand why the detected objects are outliers
  - Provide justification of the detection

General Issues: Number of Attributes

- Many anomalies are defined in terms of a single attribute
  - Height
  - Shape
  - Color
- Can be hard to find an anomaly using all attributes
  - Noisy or irrelevant attributes
  - Object is only anomalous with respect to some attributes
- However, an object may not be anomalous in any one attribute

General Issues: Number of Attributes

- Many anomaly detection techniques provide only a binary categorization
  - An object is an anomaly or it isn't
  - This is especially true of classification-based approaches
- Other approaches assign a score to all points
  - This score measures the degree to which an object is an anomaly
  - This allows objects to be ranked
- In the end, you often need a binary decision
  - Should this credit card transaction be flagged?
  - Still useful to have a score
- How many anomalies are there?

Variants of Anomaly Detection Problems

- Given a data set D, find all data points $x \in D$ with anomaly scores greater than some threshold t
- Given a data set D, find all data points $x \in D$ having the top-n largest anomaly scores
- Given a data set D, containing mostly normal (but unlabeled) data points, and a test point x, compute the anomaly score of x with respect to D

Model-Based Anomaly Detection

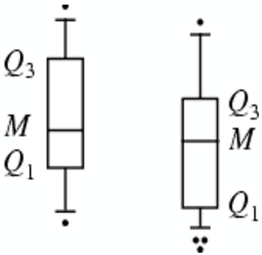
- Build a model for the data and see
- Unsupervised
  - Largely utilize clustering methods
  - Statistical methods
  - Anomalies are those points that don't fit well
  - Anomalies are those points that distort the model
- Supervised
  - Can be modeled as a classification problem
  - Special aspects to consider: anomalies are regarded as a rare class; imbalanced normal data points and abnormal points
  - Measures: recall is more meaningful
  - Need to have training data

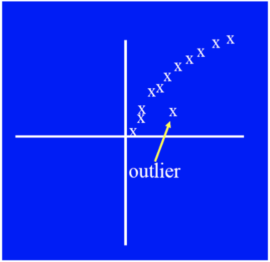
Additional Anomaly Detection Techniques

- Proximity-based
  - Anomalies are points far away from other points
  - Can detect this graphically in some cases
  - The proximity of outliers to their neighbors are different from the proximity of most other objects to their neighbors
  - Distance-based
  - Density-based
    - Low density points are outliers
- Clustering-based
  - Normal objects belong to large and dense clusters
  - Outliers belong to small or sparse clusters, or belong to no cluster

Visual Approaches

- Boxplots or scatter plots
- Limitations
  - Not automatic
  - Subjective





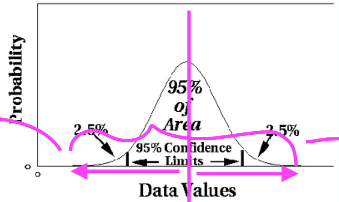
Statistical Approaches

- Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.
  - Normal objects are generated by a stochastic process, occur in regions of high probability for the stochastic model
  - Outliers occur in regions of low probability
- Approach steps
  - Learn a generative model fitting the given data
  - Identify the objects in low-probability regions of the model
- Categories
  - Parametric method (univariate, multivariate): usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
  - Nonparametric method

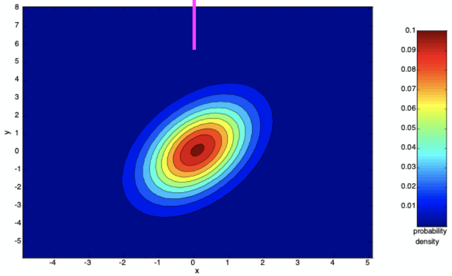
Statistical Approaches - parametric

- Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
    - Data distribution
    - Parameters of distribution (e.g., mean, variance)
    - Number of expected outliers (confidence limit)
- Issues
    - Identifying the distribution of a data set
        - Heavy tailed distribution
    - Number of attributes
    - Is the data a mixture of distributions?

Normal Distributions



One-dimensional Gaussian



Two-dimensional Gaussian

Parametric: univariate Normal Distribution

- Normal distribution, maximum likelihood estimation (MLE)
  - Standard normal distribution, $N(0, 1)$
  - Non-standard normal distribution, $N((\mu, \sigma^2)$, z-score
  - Use MLE to estimate μ, and σ^2

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

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

Parametric: univariate Normal Distribution

x<=\mu-3\sigma
x>\mu+3\sigma

- $prob(|x| \geq c) = \alpha$ for $N(0,1)$
  - Mark an object as an outlier if it is more than 3σ away from the estimated mean μ, where σ is the standard deviation ($\mu \pm 3\sigma$) region contains 99.73% of the data)
  - (c, α) pair for $N(0,1)$

c	α for $N(0,1)$
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001

Parametric: univariate Normal Distribution

(24+28.9+28.9+29+29.1+29.1+29.2+29.3+29.4)/10

((24-28.61)^2. + (28.9-28.61)^2 + ... (29.4-28.61)^2)/10

- Example
- A city' average temperature values in 10 years: 24, 28.9, 28.9, 29, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4
  - $\mu = 28.61$
  - $\sigma^2 = 2.29, \sigma = 1.51$
  - Is 24 an outlier?
    $z\text{-score} = \frac{|24-28.61|}{1.51} = 3.04$
    > 3

Grubbs' Test

- Maximum normed residual test
- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - H_0: There is no outlier in data
  - H_A: There is at least one outlier
- Grubbs' test statistic:

$$G = \frac{max(|X - \bar{X}|)}{s}$$

reject H_0 if

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/N,N-2}}{N-2+t^2_{\alpha/N,N-2}}}$$

Parametric: multivariate

- Convert the problem to a univariate outlier detection problem
- Use Mahalanobis distance from object o to its mean μ
- Use χ^2 statistic

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

  - o_i: is the value of o on the i-th dimension
  - E_i: the mean of the i-th dimension of all objects
  - n: the number of objects

Statistical-based Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t, that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A.

Statistical-based Likelihood Approach

- Data distribution, $D = (1-\lambda)M + \lambda A$
- M is a probability distribution estimated from data
- A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i)\right)\left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i)\right)$$

$$LL_t(D) = |M_t| log(1-\lambda) + \sum_{x_i \in M_t} log P_{M_t}(x_i) + |A_t| log\lambda + \sum_{x_i \in A_t} log P_{A_t}(x_i)$$

Nonparametric

- Nonparametric methods use fewer assumptions about data distribution, thus can be applicable in more scenarios
- Histogram approach
  - Construct histograms (types: equal width or equal depth, number of bins, or size of each bin)
  - Outliers: not in any bin or in bins with small size
  - Drawback: hard to decide the bin size
- Others: kernel function (more discussed in machine learning)

Strengths/Weaknesses of Statistical Approaches

- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

References

- Chapter 9: Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar