UNIVERSITY of
HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

Data Mining for Security - Overview

By
Rakesh Verma

Outline

What is Data Analytics/Data Science?

What is Security Analytics?

Case studies of attacks

What is Data Analytics?

Extracting knowledge (interesting patterns from data) using techniques from

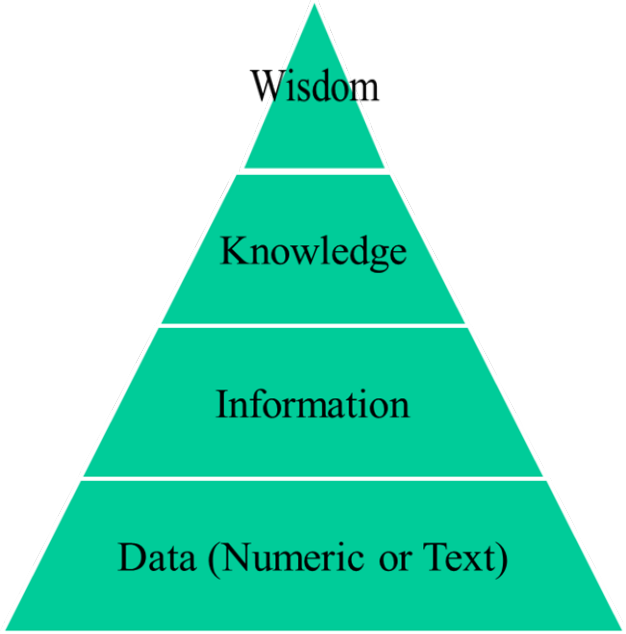
Statistics

Data Mining

Machine Learning

Text Mining/Natural Language

Processing



Interesting Patterns?

Non-trivial, implicit, previously unknown, or potentially useful

∆(Greenland sea ice area) → Monsoon rainfall in India [ToI 5/30/17]

Baby formula → Diaper

Beer → Diaper

Lower conscientiousness → Higher confidence [Asia CCS 2017]

Types of Analytics

Descriptive

Mean, median, mode, range, variance

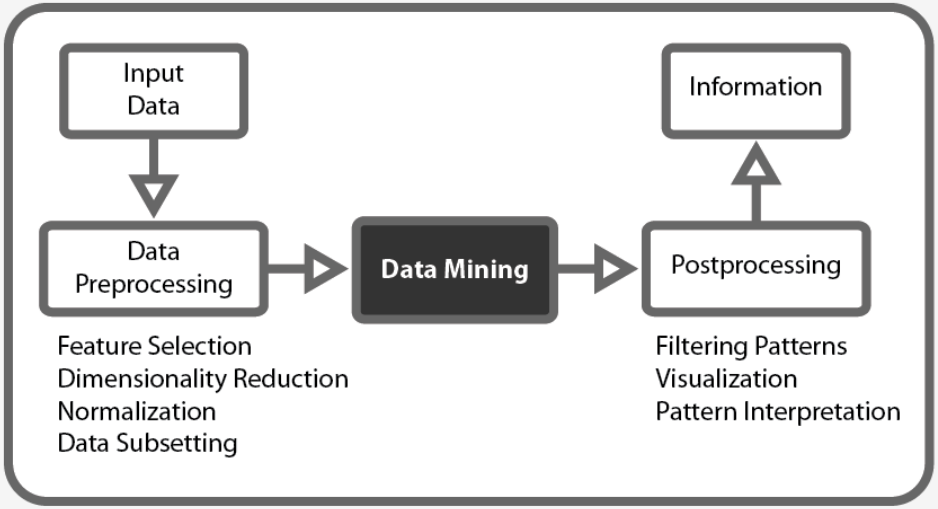
Predictive

Chance of rain is over 90% today

With this mix of investments, your returns are in the 5-6% range with 70% likelihood

Prescriptive

You should change your mix of investments



Process of knowledge discovery in databases (KDD)

Techniques for Finding Patterns

Association and Correlation Analysis

Classification

Cluster Analysis

Outlier Analysis

Techniques for Prediction

Simulations

Regression Analysis

Prescription

[www.dataskills.it]

Association and Correlation Analysis

•Frequent patterns (or frequent itemsets)

•What kinds of goods are usually bought in your Target?

•Association, correlation vs. causality

A typical association rule

Diaper → Beer [0.5%, 75%] (support, confidence)

What is the relationship between strongly associated items and strongly correlated?

•How to mine such patterns and rules efficiently in large datasets?

•How to use such patterns for classification, clustering, and other applications?

Why Association Rules? Some security applications

Malware detection [e.g. Ding et al. Computers & Security 2013]

Hypothesis: malicious behavior exhibited by system calls.

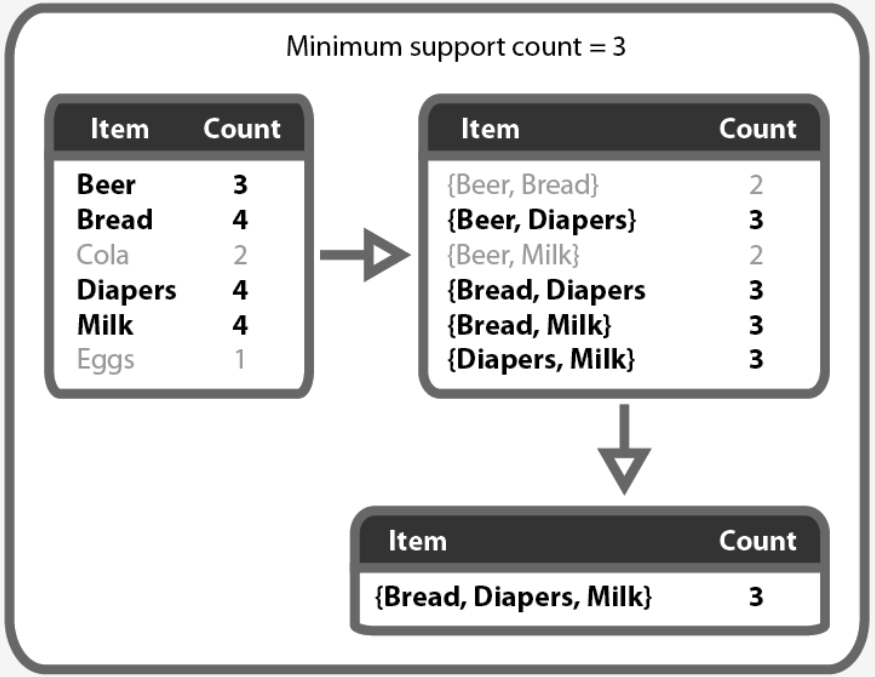
Data: API calls and frequencies: Obtained from Windows PE head file

Stepping-stone detection [e.g. Hsiao et al. Sec. and Comm. Networks 2013]

Stepping stones are intermediate hosts on the path from an hacker to a victim

Network connection records: Each transaction contains a number of pairs (s, t) where s, t are IP addresses, s – source, t - destination

ID	Items
1	Beer, Bread, Milk, Diapers
2	Beer, Diapers, Milk
3	Beer, Bread, Milk, Cola, Diapers
4	Bread, Cola, Milk, Diapers
5	Bread, Eggs



Frequent itemset generation using the Apriori algorithm

Classification

Classification and label prediction

Construct models (functions) based on some training examples

Describe and distinguish classes or concepts for future prediction

E.g., classify countries based on (climate), or classify cars based on (gas mileage)

Predict some unknown class labels

Typical methods

Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …

Classification

Typical application

Credit/loan approval

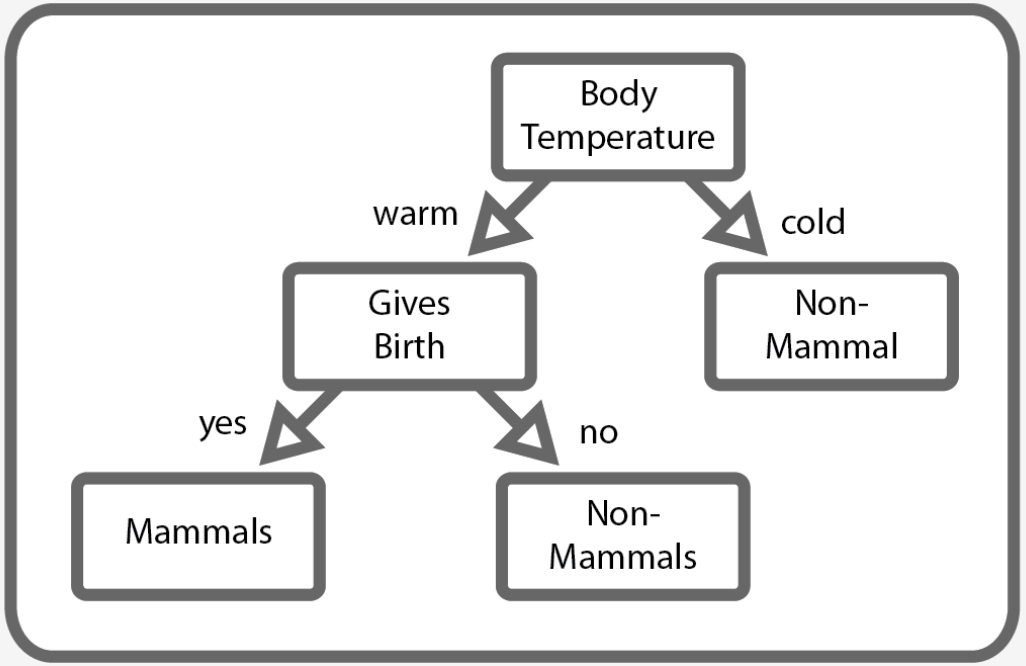
Medical diagnosis: if a tumor is cancerous or benign

Fraud detection: if a transaction is fraudulent

Web page categorization: which category it is

Malware detection: is a piece of software malicious

Malicious URL detection, Intrusion detection, …



Decision tree for mammal classification problem

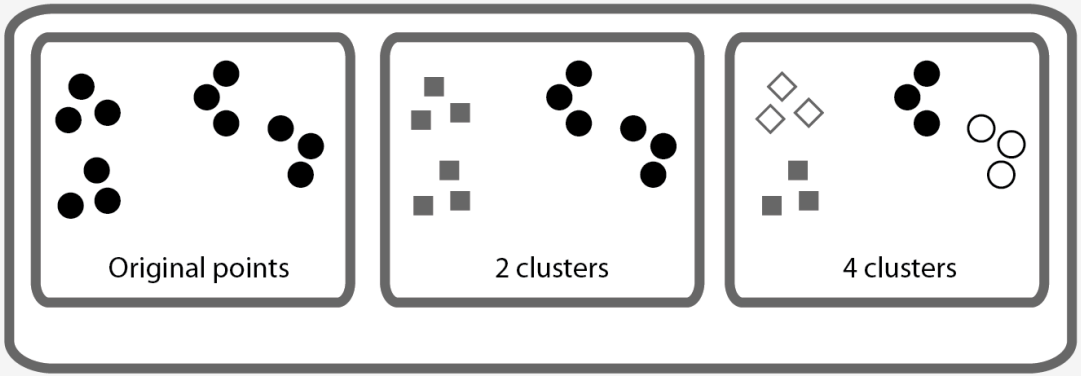
Cluster Analysis

Unsupervised learning (i.e., Class label is unknown)

Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns

Principle: Maximizing intra-class similarity & minimizing interclass similarity

Many methods and applications



Different ways of clustering the same set of points

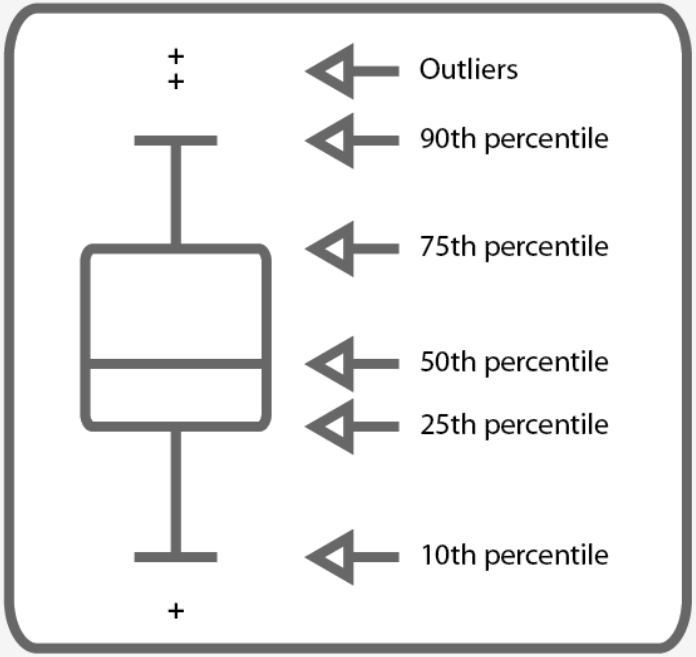
Outlier Analysis

Outlier: A data object that does not comply with the general behavior of the data

Noise or exception? – One person's garbage could be another person's treasure

Methods: byproduct of clustering, or regression analysis, …

Useful in fraud detection, rare events analysis


Outliers
90th percentile
75th percentile
50th percentile
25th percentile
10th percentile

Boxplot

Types Of Data

Data to be mined

Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

Knowledge to be mined (or: Data mining functions)

Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.

Descriptive vs. predictive data mining

Multiple/integrated functions and mining at multiple levels

Techniques utilized

Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

Applications adapted

Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Types of Data

Database-oriented data sets and applications

Relational database, data warehouse, transactional database

Object-relational databases, Heterogeneous databases and legacy databases

Advanced data sets and advanced applications

Data streams and sensor data

Time-series data, temporal data, sequence data (incl. bio-sequences)

Structure data, graphs, social networks and information networks

Spatial data and spatiotemporal data

Multimedia database

Text databases

The World-Wide Web

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

Sequence, trend and evolution analysis

Trend, time-series, and deviation analysis: e.g., regression and value prediction

Sequential pattern mining

e.g., first buy digital camera, then buy large SD memory cards

Periodicity analysis

Motifs and biological sequence analysis

Approximate and consecutive motifs

Similarity-based analysis

Mining data streams

Ordered, time-varying, potentially infinite, data streams

Types of Data Sets

Record

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vecto
- Transaction data

Graph and network

- World Wide Web
- Social or information networks
- Molecular Structures

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Ordered

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

Spatial, image and multimedia:

- Spatial data: maps
- Image data:
- Video data:

What is Security Analytics?

Adaptation (not direct application) of tools and techniques from

Statistics

Data Mining

Machine Learning

Natural Language Processing

to challenges in cybersecurity

Why Adaptation?

Active adversary

Time scale of attacks

Availability/diversity of data

Unbalanced datasets, costs of misclassification

Base rate fallacy

[More at: Verma et al., IEEE Security and Privacy Mag. Nov/Dec 2015]

Some examples

Naïve Bayes, Support Vector Machines, Neural networks, Decision Trees, …, used for:

Intrusion detection

Malware detection

Spam

Phishing, Spear phishing (anomaly detection),

…

UNIVERSITY of
HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

Data Mining

Rakesh Verma

This is Information

00010000 11110000
00011101 11110000
00010110 11110000
00011011 11110000
00110000 10101010
00101000 10101010
00100000 10101010
01110111 00001111
01010111 00001111
01111111 00001111

This is knowledge

00010000 **1111**0000
00011101 **1111**0000
00010110 **1111**0000
00011011 **1111**0000
00110**000** 10101010
00101**000** 10101010
00100**000** 10101010
01110**111** 00001111
01010**111** 00001111
01111**111** 00001111

Datamining

Data mining is the act of extracting interesting non-trivial, novel, and useful patterns (knowledge) in large amounts of data (information).

These patterns are often very difficult to discover for a human being, but very easy for a machine to discover with the proper algorithms.

Why we need Data Mining

We live in a data-rich society.

- Everyone is connected to the internet across multiple apps.
- Every tweet, blog post, and pageview you make is data.
- There is over 1 exabyte (1000^6 bytes) of traffic on the internet every day.

We also live in a knowledge-poor society.

- When looking at all this data, we don't know what to do with it.
- Or we know what to do with it, but not how.

Classes of Data Mining Tasks

Prediction

Given some context from currently held data, predict unknown or future information or events.

Examples: Stock prices, spam detection, fraud detection, weather forecasts, patient diagnosis

Description

Extract patterns in the data that represent higher-level knowledge.

Examples: Automatic speech recognition, image identification, edge detection, botnet life cycle categorization (Guo CCS2008)

Types of Data Mining Tasks

Associations

What types of data correlate with each other more regularly.

Recommended videos on Youtube, "People who bought this also bought" Amazon

Classification

Inferring the identities of this data.

Snapshot Serengeti, Facial Recognition, Flooding Prediction, Phishing Website Detection

Cluster Analysis

Grouping data together into similarity categories

Botnets Network Traffic, Phishing Campaing, Marketting, Classification without prior knowledge

Outlier Analysis

Analysing the data points that seem out of place

Noise removal, false-positive detection, fraud detection, DDoS attack detection

Association

Frequent Itemsets: Events that appear with each other frequently.

Association Rules

- People who make their friend list public on social media tend to fall for phish
- People who buy diapers tend to buy beer as well.
- Unexpected patterns like the above are plentiful and significant
  - Buying beer doesn't cause buying diapers, however. Correlation not causation.

Typical Algorithms

- Apriori algorithm

Classification

Class Label Prediction

- Supervised method using training data.
- Separates data instances into classes or categories
- Infers true classification/identity of instances.

Typical Algorithms

- Decision Trees
- Neural Networks
- Bayesian classifiers
- Logistic regression
- Support vector machines

Clusters

Cluster Grouping Instances

Unsupervised method used without training data,

Groups instances together into groups based on similarity or proximity.

Can be used in situations where classification would be used if a training set existed.

Cluster Analysis has numerous algorithms to solving this problem

Clustering is a very broad umbrella term.

Outlier Analysis

Exclude degenerate instances

- Instances that are not stereotypical of their classes can negatively impact experiments
- Removing problem instances in training data improves results.
- Identifying exception cases could lead to the discovery of new patterns.

Typical Algorithms

- Clustering
- Regression

Dataset Types

- Relational Databases
  - Relational-style row-by-column tables where each row is a data instance
- Sensor Data and Time Series
  - Thermometers, Accelerometers, Timers, Sound
  - Time series are data points collected over time with timestamps for each point
    - Video also falls into this category
- Spatial Data: Geographical map data, 3D positioning data, images, video.
- Graph and Network Data: Social Networks, Tree Structured Data, Graphs
- Internet: Website DOMs

Dataset Types

- Tabular
  - Matrices of row-vectors, columns are features
  - Relational Tables
  - Bags or sets
- Graph and Network
  - Social Media
  - The Internet
- Sequential
  - Sound, Video
  - DNA
- Spatial
  - Images
  - Maps

Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Properties of Datasets

- Dimensionality
  - How many different parameters does each instance in the dataset have?
- Sparsity
  - How often are parameters missing for each instance?
  - Is their absence important or meaningless?
- Resolution
  - How finely measured is the dataset?
  - How big are the units?
  - What impact does this have on the results?
- Distribution
  - How similar are all the instances to each other?
  - How different are they from each other?

Data Objects

Data sets are made up of data objects.

A **data object** represents an entity.

Examples:

sales database: customers, store items, sales

medical database: patients, treatments

university database: students, professors, courses

Also called *samples , examples, instances, data points, objects, tuples*.

Data objects are described by **attributes**.

Database rows -> data objects; columns ->attributes.

Attributes

Attribute (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.

E.g., customer _ID, name, address

Types:

Nominal

Binary

Numeric: quantitative

Interval-scaled

Ratio-scaled

Attribute Types

Nominal: categories, states, or "names of things"

Hair_color = {*auburn, black, blond, brown, grey, red, white*}

marital status, occupation, ID numbers, zip codes

Binary

Nominal attribute with only 2 states (0 and 1)

Symmetric binary: both outcomes equally important

e.g., gender

Asymmetric binary: outcomes not equally important.

e.g., medical test (positive vs. negative)

Convention: assign 1 to most important outcome (e.g., HIV positive)

Ordinal

Values have a meaningful order (ranking) but magnitude between successive values is not known.

Size = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

Quantity (integer or real-valued)

Interval

- Measured on a scale of **equal-sized units**
- Values have order
  - E.g., *temperature in C˚ or F˚, calendar dates*
- No true zero-point

Ratio

- Inherent zero-point
- We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
  - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

Discrete Attribute

Has only a finite or countably infinite set of values

E.g., zip codes, profession, or the set of words in a collection of documents

Sometimes, represented as integer variables

Note: Binary attributes are a special case of discrete attributes

Continuous Attribute

Has real numbers as attribute values

E.g., temperature, height, or weight

Practically, real values can only be measured and represented using a finite number of digits

Continuous attributes are typically represented as floating-point variables

Attribute Type		Description	Examples
Categorical (Qualitative)	Nominal	Descriptive info (= , ≠)	Eye color
	Ordinal	Order, ranks (< , >)	Small, Medium, Large
Numeric (Quantitative)	Interval	Measurements (+ , -)	Calendar Dates
	Ratio	Proportionality (* , /)	Age, Length

Different attribute types

Data Quality: Why Preprocess the Data?

Measures for data quality: A multidimensional view

Accuracy: correct or wrong, accurate or not

Completeness: not recorded, unavailable, …

Consistency: some modified but some not, dangling, …

Timeliness: timely update?

Believability: how trustable the data are correct?

Interpretability: how easily the data can be understood?

Tasks of Data Preprocessing

Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

Integration of multiple databases, data cubes, or files

Data reduction

Dimensionality reduction

Numerosity reduction

Data compression

Data transformation and data discretization

Normalization

Concept hierarchy generation

Instances

Datasets are composed of instances.

Also known as samples, examples, data points.

Each instance represents one entity.

One website in a dataset of websites.

One DNA sequence in a genome dataset.

Each instance is described by a number of features.

Each website has a <body> tag with content in it.

Every DNA sequence has nucleotide pairs.

Features

Features (sometimes called attributes, dimensions) describe the instance.

Feature representations often fall into three types:

- Nominal
  - Names, categories, words.
- Binary
  - Instance has the feature or doesn't.
- Numeric
  - Instance has this score or count of a feature.
    - Interval-Scaled
      - Equally-sized units with an order.
      - Temperature in Celsius, Position relative to Sea-level.
    - Ratio-Scaled
      - Have an inherent zero-point, as well as an order.
      - Temperature in Kelvin, Pageview counts.

Discrete Features vs. Continuous Features

Discrete

A finite or countably infinite number of values.

Integer-valued features, sets of elements, binary attributes.

Even with order, there's not always a concept of "in-between" for two values.

Continuous

Can take any real or rational number value.

Precise temperatures, time taken, physical size.

In practice, real numbers have limited precision due to computational limits.

There is a concept of values that are in-between.

Data Quality

Accuracy

How correct were the measurements? Are there any wrong measurements?

Believability

How trustworthy is the dataset? Is it artificially created or true to actual observations?

Completeness

How many missing values? How many instances were collected? Is that enough to be characteristic of the problem?

Consistency

Which instances are up-to-date? Are some of the instances encoded in a different convention?

Interpretability

How easy is it to utilize the dataset the way it's encoded? Is there a lot of overhead in using it?

Preprocessing - Improving the Quality of Data

Data Cleaning

Noise Reduction, Outlier Removal, Missing Value Replacement

Data Integration or Data Coalescing

Combining multiple datasets or files of the same kind of data together.

Data Reduction

Reducing the number of features.

Accounting for too many redundant or similar instances.

Reducing overall data size through compression.

Data Transformation

Data normalization.

Converting from row-vectors to bags or sets.

Altering the topology of the data.

References

1. BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection