Descriptive analysis, preprocessing, visualisation...

Jaak Vilo
2017 Spring

Characterise data

use Big_University_DB

mine characteristics as "Science_Students" in relevance to
name, gender, major, birth_date, residence, phone#, gpa
from student
where status in graduate

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process
- Data in the real world is dirty
- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=""
- noisy: containing errors or outliers
  - e.g., Salary="-10"
- inconsistent: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"
- e.g., discrepancy between duplicate records

Why Data Preprocessing?

- Data in the real world is dirty
- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- e.g., occupation=""
- noisy: containing errors or outliers
- e.g., Salary="-10"
- inconsistent: containing discrepancies in codes or names
- e.g., Age="42" Birthday="03/07/1997"
- e.g., Was rating "1,2,3", now rating "A, B, C"
- e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
  - Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
- Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse
- Garbage in, garbage out
High quality data requirements

- High-quality data needs to pass a set of quality criteria. Those include:
  - **Accuracy**: an aggregated value over the criteria of integrity, consistency, and density
  - **Integrity**: an aggregated value over the criteria of completeness and validity
  - **Completeness**: achieved by correcting data containing anomalies
  - **Validity**: approximated by the amount of data satisfying integrity constraints
  - **Consistency**: concerns contradictions and syntactical anomalies
  - **Uniformity**: directly related to irregularities and in compliance with the set 'unit of measure'
  - **Density**: the quotient of missing values in the data and the number of total values ought to be known

http://en.wikipedia.org/wiki/Data_cleansing

February 22, 2017 Data Mining: Concepts and Techniques

---

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - **Accuracy**
  - **Completeness**
  - **Consistency**
  - **Timeliness**
  - **Believability**
  - **Value added**
  - **Interpretability**
  - **Accessibility**

- Broad categories:
  - Intrinsic, contextual, representational, and accessibility

February 22, 2017 Data Mining: Concepts and Techniques

---

Major Tasks in 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 transformation**
  - Normalization and aggregation
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data discretization**
  - Part of data reduction but with particular importance, especially for numerical data

February 22, 2017 Data Mining: Concepts and Techniques

---

Forms of Data Preprocessing

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

February 22, 2017 Data Mining: Concepts and Techniques

---

Chapter 2: Data Preprocessing

- Why preprocess the data?
- **Descriptive data summarization**
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

February 22, 2017 Data Mining: Concepts and Techniques

---

| 0.48 0.03 0.06 0.05 0.43 0.19 0.16 0.35 0.25 0.07 |
| 0.29 0.14 0.96 0.02 0.11 0.22 0.80 0.05 0.54 0.36 |
| 0.23 0.28 0.02 0.10 0.48 0.31 0.36 0.21 0.33 0.45 |
| 0.64 0.04 0.48 0.56 0.16 0.58 0.33 0.11 0.42 0.06 |
| 0.00 0.23 0.24 0.00 0.54 0.02 0.26 0.20 0.18 0.01 |
| 0.17 0.17 0.04 0.97 0.25 0.04 0.34 0.01 0.50 0.15 |
| 0.43 0.05 0.50 0.16 0.52 0.82 0.23 0.09 0.02 0.21 |
| 0.13 0.17 0.33 0.26 0.00 0.33 0.57 0.43 0.09 0.43 |
| 0.24 0.08 0.08 0.54 0.08 0.02 0.01 0.01 0.35 0.62 |
| 0.10 0.03 0.14 0.78 0.30 0.07 0.08 0.48 0.48 0.57 |

Jaak Vilo and other authors

UT: Data Mining 2009

February 22, 2017 Data Mining: Concepts and Techniques
Mining Data Descriptive Characteristics

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for grouped data):
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: \( \text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median}) \)
Histograms and Probability Density Functions

- Probability Density Functions
  - Total area under curve integrates to 1
- Frequency Histograms
  - Simple interpretation
  - Can't be directly related to probabilities or density functions
- Relative Frequency Histograms
  - Divide counts by total number of observations
  - Y-axis is counts
  - Simple interpretation
  - Can't be directly related to density functions
  - Bar heights can be >= 1 but each range is 1 divided by bar width
- Density Histograms
  - Divide counts by (total number of observations X bar width)
  - Y-axis is density values
  - Simple interpretation
  - Can be directly related to density functions
  - Bar area can be >= 1

Jaak Vilo and other authors
UT: Data Mining 2009
http://www.geog.ucsb.edu/~joel/g210_w07/lecture_notes/lect04/oh07_04_1.html

The data are (the log of) wing spans of aircraft built in from 1956 - 1984.
2-dimensional data: dot-plot

Histogram vs kernel density

• properties of histograms with these two examples:
  – they are not smooth
  – depend on end points of bins
  – depend on width of bins

• We can alleviate the first two problems by using kernel density estimators.
• To remove the dependence on the end points of the bins, we centre each of the blocks at each data point rather than fixing the end points of the blocks.

• Blocks - it is still discontinuous as we have used a discontinuous kernel as our building block
• If we use a smooth kernel for our building block, then we will have a smooth density estimate.
• It’s important to choose the most appropriate bandwidth as a value that is too small or too large is not useful.
• If we use a normal (Gaussian) kernel with bandwidth or standard deviation of 0.1 (which has area 1/12 under the each curve) then the kernel density estimate is said to undersmoothed as the bandwidth is too small in the figure below.
• It appears that there are 4 modes in this density - some of these are surely artifacts of the data.

• Choose optimal bandwidth
  –Methods to estimate it

  • AMISE = Asymptotic Mean Integrated Squared Error
  • optimal bandwidth = arg min AMISE

• The optimal value of the bandwidth for our dataset is about 0.25.
• From the optimally smoothed kernel density estimate, there are two modes. As these are the log of aircraft wing span, it means that there were a group of smaller, lighter planes built, and these are clustered around 2.5 (which is about 12 m).
• Whereas the larger planes, maybe using jet engines as these used on a commercial scale from about the 1960s, are grouped around 3.5 (about 33 m).
• The properties of kernel density estimators are, as compared to histograms:
  – smooth
  – no end points
  – depend on bandwidth

Kernel Density estimation - links

• Wikipedia:
  http://en.wikipedia.org/wiki/Kernel_density_estimation
• Ricardo Gutierrez-Osuna
  http://research.cs.tamu.edu/prism/lectures/pr/pr_l7.pdf
• Tutorial and Java applet for testing:
• Simple interactive web demo at U. Tartu:
  – http://kde.tume-maailm.pri.ee/

Definition

If \( x_1, x_2, \ldots, x_n \sim f \) is an independent and identically-distributed sample of a random variable, then the kernel density approximation of its probability density function is
\[
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right)
\]
where \( K \) is some kernel and \( h \) is a smoothing parameter called the bandwidth. Quite often \( K \) is taken to be a standard Gaussian function with mean zero and variance 1. Thus the variance is controlled indirectly through the parameter \( h \):
\[
K \left( \frac{x - x_i}{h} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}.
\]

R – example (due K. Tretjakov)

\[
d = c(1,2,2,2,2,2,2,2,2,3,2,3,4,5,4,3,2,3,4,4,5,6,7);
kernelsmooth <- function(data, sigma, x) {
  result = 0;
  for (d in data) {
    result = result + exp(-d^2/2/sigma^2);
  }
  result/sqrt(2*pi)/sigma;
}
x = seq(min(d), max(d), by=0.1);
y = sapply(x, function(x) {kernelsmooth(d, 1, x)});
hist(d);
lines(x,y);

Aggregation, analysis and visualization of geodata

- http://sightsmap.com
- Several large crowd-sourced datasets:
  - The whole Panoramio photobank used by Google maps
  - The whole Wikipedia, geotags and wikipedia article logs
  - Foursquare
  - ... More
- Aggregate data, calculate popularities of places, calculate type tags
- Translate and categorize titles and descriptions to get types
- Improve aggregation algorithms by learning
- Visualize heatmaps, type tags, aggregated sources

By: Tanel Tammet
Jaak Vilo and other authors

UT: Data Mining 2009

http://176.32.89.45/~hideaki/res/kernel.html
MDL Histogram Density Estimation

Jaak Vilo and other authors

UT: Data Mining 2009

Abstract

We regard histogram density estimation as a model selection problem. Our approach to the problem is based on the so-called MDL principle, which can be applied to both fixed- and data-clustering density estimation, image-describing and model selection in general. MDL-based model selection is formulated via the normalized maximum likelihood (NML) approach. The NML density has optimal properties. We show how this knowl-

dge can be used for selecting the optimal bin count. Three super-classes are, however, specified: it has been shown that there are only few models that have small superset and large counts, respectively. These models are only good for describing roughly unknown data. If the data distribution is strongly non-uniform, the bin count must necessarily be high in one region to capture the details of the high density portion of the distribution. On the other hand, the large model class has to become less dense in the low density region.

To avoid the problems of uniform histograms we must allow the bin count to be variable width. For these super-

classes it is necessary to find the optimal set of density parameters.

More links on R and kernel density

- http://sekhon.berkeley.edu/stats/html/density.html
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: \( \text{IQR} = Q_3 - Q_1 \)
- Five number summary: min, Q1, M, Q3, max
- Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
- Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (sample: \( s \), population: \( \sigma \))
  - Variance: (algebraic, scalable computation)
    \[
    s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left( \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \mu^2
    \]
  - Standard deviation \( s \) (or \( \sigma \)) is the square root of variance \( s^2 \)(or \( \sigma^2 \))

Boxplot Analysis

- Five-number summary of a distribution:
  - Minimum, Q1, M, Q3, Maximum
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
Box Plots

- http://informationandvisualization.de/blog/box-plot

1.5 x IQR – Inter Quartile range

10.7 cm Solar Flux

Observed 10.7 cm Solar Flux

Min. Max
1% 99%
Median
Mean
square
box
25%-75%
Visualization of Data Dispersion: Boxplot Analysis

A Box Plot can show the difference in variance between replicates

Variance regularization can remove the bias

Violin plot

Violin plot – R

vertical kernel density

http://gallery.renthusiasts.com/graph/Violin_plot,43

Combining plots...
Properties of Normal Distribution Curve

- 99.7% of the data are within 3 standard deviations of the mean.
- 95% within 2 standard deviations.
- 68% within 1 standard deviation.

Example of plots-containing article:
- [http://www.kgs.ku.edu/Magellan/WaterLevels/CD/Reports/OFR04_57/rep00.htm](http://www.kgs.ku.edu/Magellan/WaterLevels/CD/Reports/OFR04_57/rep00.htm)

Quantile-Quantile (q-q) Plots

- [http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html](http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html)

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
These 2 batches do not appear to have come from populations with a common distribution.

The batch 1 values are significantly higher than the corresponding batch 2 values.

The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

The advantages of the q-q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
Parametric modeling usually involves making assumptions about the shape of data, or the shape of residuals from a regression fit. Verifying such assumptions can take many forms, but an exploration of the shape using histograms and q-q plots is very effective. **The q-q plot does not have any design parameters such as the number of bins for a histogram.**

Kemmeren et al. (Mol. Cell, 2002)

**Randomized expression data**
- Yeast 2-hybrid studies
- Known (literature) PPI

MPK1 YLR350w
SNF4 YCL046W
SNF7 YGR122W.

Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

Pearson Correlation

\[
\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2} \sqrt{E(Y^2) - E(Y)^2}}
\]

\[
\tau = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}
\]
Not Correlated Data

Bias vs Variance

Numerical summary?

Anscombe’s quartet

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of $x$ in each case</td>
<td>11 exact</td>
</tr>
<tr>
<td>Variance of $x$ in each case</td>
<td>10 exact</td>
</tr>
<tr>
<td>Mean of $y$ in each case</td>
<td>7.50 (to 2 d.p.)</td>
</tr>
<tr>
<td>Variance of $y$ in each case</td>
<td>3.75 (to 2 d.p.)</td>
</tr>
<tr>
<td>Correlation between $x$ and $y$ in each case</td>
<td>0.816 (to 3 d.p.)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 d.p. and 3 d.p.) resp.</td>
</tr>
</tbody>
</table>
Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence.
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression.

Graphic Displays of Basic Statistical Descriptions

- Histogram: (shown before)
- Boxplot: (covered before)
- Quantile plot: each value $x_i$ is paired with $f_i$ indicating that approximately $100 f_i\%$ of data are $\leq x_i$.
- Quantile-quantile (Q-Q) plot: graphs the quantiles of one univariate distribution against the corresponding quantiles of another.
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane.
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence.

Chapter 2: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
  - Missing data may be due to
    - equipment malfunction
    - inconsistent with other recorded data and thus deleted
    - data not entered due to misunderstanding
    - certain data may not be considered important at the time of entry
    - not register history or changes of the data
    - Missing data may need to be inferred.
How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant: e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

K-NN impute

- K nearest neighbours imputation
- Find K neighbours on available data points
- Estimate the missing value
- (Hastie, Tibshirani, Troyanskaya, ... Stanford 1999-2001)

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B–A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky
Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Regression

\[ y = x + 1 \]

Cluster Analysis

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter’s Wheels)

Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
  - Schema integration: e.g., A.cust-id = B.cust-id
  - Integrate metadata from different sources
  - Entity identification problem:
    - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
  - Detecting and resolving data value conflicts
    - For the same real world entity, attribute values from different sources are different
    - Possible reasons: different representations, different scales, e.g., metric vs. British units
Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - **Object identification**: The same attribute or object may have different names in different databases
  - **Derivable data**: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by **correlation analysis**
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

### Normalisation

- Making data comparable...

### Data preprocessing

Record linkage is highly sensitive to the quality of the data being linked, so all data sets under consideration (particularly their key identifier fields) should ideally undergo a data quality assessment prior to record linkage. Many key identifiers for the same entity can be presented quite differently between (and even within) data sets, which can greatly complicate record linkage unless understood ahead of time. For example, key identifiers for a man named William J. Smith might appear in three different data sets as so:

<table>
<thead>
<tr>
<th>Data set</th>
<th>Name</th>
<th>Date of birth</th>
<th>City of residence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 1</td>
<td>William J. Smith</td>
<td>1/07/73</td>
<td>Berkeley, California</td>
</tr>
<tr>
<td>Data set 2</td>
<td>Smith, W. J.</td>
<td>1/07/73</td>
<td>Berkeley, CA.</td>
</tr>
<tr>
<td>Data set 3</td>
<td>J. Smith</td>
<td>Jan 2, 1973</td>
<td>Calif.</td>
</tr>
</tbody>
</table>

In this example, the different formatting styles lead to records that look different but in fact all refer to the same entity with the same logical identifier values. Most, if not all, record linkage strategies would result in more accurate linkage if these values were first normalized or standardized into a consistent format (e.g., all names are “Surname, Given-name”, all dates are “YYYY/MM/DD”, and all cities are “Name, 2-letter state abbreviation”). Standardization can be accomplished through simple rule-based transformations or more complex procedures such as lexicon-based translation and probabilistic hidden Markov models.  

### Normalisation

- **Making data comparable**...
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Elements of microarray statistics

Reference

Test

$M = \log_2 R - \log_2 G = \log_2 (R/G)$

$A = \frac{1}{2} (\log_2 R + \log_2 G)$

Normalisation
can be used to transform data

Data Reduction Strategies

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation:
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Discretization

- Three types of attributes:
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Discretization and Concept Hierarchy

- Discretization
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

E.g.

- 0-18 y: 1580
- 19-65 y: 8394
- 66-105: 2700

Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, "natural" intervals.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Example of 3-4-5 Rule
Example

-351,976.00 4,700,896.50
MN = -351,976.00
MAX = 4,700,896.50
LOW = 5th percentile -159,876
HIGH = 95th percentile 1,038,761
msd = 1,000,000 (most significant digit)
LOW = 1,000,000 (round down) HIGH = 2,000,000 (round up)

Adjust with real MIN and MAX
1. (-400,000 .. 0]
2. (0 .. 1,000,000]
3. (1,000,000 .. 2,000,000]

Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - (Urbana, Champaign, Chicago) < Illinois
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: (street, city, state, country)

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year

Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Discriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research
References

- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999