Scientific Data Analysis: Data Preprocessing

Jarek Szlichta http://data.science.uoit.ca/

Data, Data Everywhere...

Open data



Business Data



Web Data



Available at different





Big Data to Data Science



Data Scientist: The Sexiest Job of the 21st Century

Harvard Business Review Oct. 2012

(c) 2012 Biocomicals by Dr. Alper Uzon

Can we take data science out of the realm of geekdom and make it a trusted, respected profession?

Poor Data Quality

- Data quality is an increasingly pervasive problem for organizations
- Operational inefficiencies and mistakes
 - Major Canadian bank faxed sensitive customer data to US junk yard (incorrect extra digit in fax number)
- Waste of money and time [Gartner Research Report `09]
 - Companies lose, on average, \$8M annually due to poor quality data
 - 25% of critical data in the Fortune 1,000 companies is inaccurate

Data Preprocessing

- Getting to Know Your Data
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Types of Data Sets

- Record
 - Relational records
 - text documents
- Graph and network
 - World Wide Web
 - Social networks
 - Molecular structures
- 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

	team	coach	pla y	ball	score	game	wi n	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

Data Objects

- Datasets are made up of data objects.
- A data object represents a tuple.
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Tuples are described by **attributes**.

Attributes

- Attribute: 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, zip codes
- Binary
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric binary: both outcomes equally important
 - e.g., gender
 - <u>Asymmetric binary</u>: 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
- Ratio
 - 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

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 floatingpoint variables

Measuring the Central Tendency

- Mean (algebraic measure):
 - Weighted arithmetic mean:
- Median:
 - Middle value if odd number of values, or average of the middle two values otherwise
 - 1, 3, 3, 6, 7, 8, 9; Median = 6

Mode

- Value that occurs most frequently in the data
- Empirical formula:

 $mean - mode = 3 \times (mean - median)$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Symmetric vs. Skewed Data

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







Data Quality: Why Preprocess the Data?

- Measures for data quality:
 - Accuracy: correct or wrong
 - Completeness: not recorded, unavailable, ...
 - Timeliness: timely update?

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 reduction

- Dimensionality reduction
- Data compression
- Data transformation and data discretization
 - Normalization
 - Binning

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., human or computer error, transmission error
 - <u>incomplete</u>: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., Occupation="" (missing data)
 - <u>noisy</u>: containing noise, errors, or outliers
 - e.g., Salary="-10" (an error)
 - inconsistent: containing discrepancies in codes or names, e.g.,
 - Age="42", Birthday="03/07/2010"
 - Intentional (e.g., disguised missing data)
 - Jan. 1 as everyone's birthday?

Incomplete (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
- Missing data may need to be inferred!

How to Handle Missing Data?

- Ignore the tuple:—not effective when the % 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 (e.g. clustering): smarter
 - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- Incorrect attribute values may be due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention
- Other data problems which require data cleaning
 - duplicate records
 - 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)

Data Integration

- Data integration:
 - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id = B.cust-#
 - 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 and covariance analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

X² (chi-square) test

$$\chi^{2} = \sum \frac{(Observed - Expected)^{2}}{Expected}$$

- The larger the X² value, the more likely the variables are related
- The cells that contribute the most to the X² value are those whose actual count is very different from the expected count
- Correlation does not imply causality
 - # of hospitals and # of car-theft in a city are correlated
 - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- For example the expected value for the cell
 - (play chess, like science fiction) = (300 x 450) / 1500 = 90
- X² (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^{2} = \frac{(250 - 90)^{2}}{90} + \frac{(50 - 210)^{2}}{210} + \frac{(200 - 360)^{2}}{360} + \frac{(1000 - 840)^{2}}{840} = 507.93$$

 It shows that like_science_fiction and play_chess are correlated in the group

Data Reduction Strategies

- Data reduction: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
 - Dimensionality reduction, e.g., remove unimportant attributes
 - Feature subset selection, feature creation
 - Numerosity reduction (some simply call it: Data Reduction)
 - Regression
 - Clustering, sampling
 - Data compression

Attribute Subset Selection

- Redundant attributes
 - Duplicate much or all of the information contained in one or more other attributes
 - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
 - Contain no information that is useful for the data mining task at hand
 - E.g., students' ID is often irrelevant to the task of predicting students' GPA

Parametric Data Reduction: Regression

Linear regression

- Data modeled to fit a straight line
- Often uses the least-square method to fit the line

Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a *dependent variable* and of one or more *independent variables*
- The parameters are estimated so as to give a "best fit" of the data
- Most commonly the best fit is evaluated by using the *least squares method*, but other criteria have also been used



 Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regress Analysis Models

- Linear regression: Y = w X + b
 - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
 - Using the least squares criterion to the known values of $Y_1, Y_2, ..., X_1, X_2$,

••••

Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- There are many choices of clustering definitions and clustering algorithms

Sampling

- Sampling: obtaining a small sample s to represent the whole data set N
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
 - Develop adaptive sampling methods, e.g., stratified sampling.

Types of Sampling

Simple random sampling

- There is an equal probability of selecting any particular item
- Sampling without replacement
 - Once an object is selected, it is removed from the population

Sampling with replacement

A selected object is not removed from the population

Stratified sampling:

- Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
- Used in conjunction with skewed data

Sampling: With or without Replacement



Sampling: Cluster or Stratified Sampling

Raw Data

Cluster/Stratified Sample





Normalization

Min-max normalization: to [new_min_A, new_max_A]

$$v' = \frac{v - min_{A}}{max_{A} - min_{A}} (new max_{A} - new min_{A}) + new min_{A}$$

 Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to

$$\frac{73,600-12,000}{98,000-12,000}(1.0-0) + 0 = 0.716$$

Discretization

- Discretization: Divide the range of a continuous attribute into intervals
 - Interval labels can then be used to replace actual data values
 - Reduce data size by discretization
 - Discretization can be performed recursively on an attribute

Data Discretization Methods

- Typical methods: All the methods can be applied recursively
 - Binning
 - Clustering analysis

Simple Discretization: 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 (equl-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

Concept Hierarchy Generation

- Concept hierarchy organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate <u>drilling and rolling</u> in data warehouses to view data in multiple granularity
- 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 *youth, adult,* or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
 - street < city < state < country</p>
- 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



15 distinct values

365 distinct values

3567 distinct values

674,339 distinct values

Summary

- Data quality: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- Data integration from multiple sources:
 - Entity identification problem
 - Remove redundancies
 - Detect inconsistencies
- Data reduction
 - Dimensionality reduction
 - Data compression
- Data transformation and data discretization
 - Normalization
 - Binning

Reading List

Recommended

- Review Slides!
- Book: Jiawei Han, Micheline Kamber and Jian Pei, Data Mining -Concepts and Techniques, Morgan Kaufmann, Third Edition, 2011 (or 2nd edition)
 - <u>http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining BOOK.pdf</u>
 - Chapter: 2