Page no.: ____/
Date: __/__/____

Data Cleaning

Real world data → tend to be incomplete, noisy and inconsistancy.

→ Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistancies in data.

• incomplete • noisy • inconsistent • intentional

◊ missing values:

Data is not always available.

eg: customer_income = " ",

missing values due to
- equipment malfunction
- inconsistent thus deleted.
- not entered due to misunderstanding
- not imp while entry

methods:

1) Ignore the tuple: usually done when class label is missing.
   → not effective
   → poor when % of missing values per attribute varies considerably.
   → By ignoring we do not make use of remaining attribute's values in the tuple.

2) Fill in the missing value manually: Time-consuming & may not be feasible given a large data set in many missing values.

3) Use a global constant to fill in missing values: "Unknow" or $-\infty$
   This method is simple not fullproof

Page no.:
Date: / /

4) Use a measure of central tendency for the attribute to fill in the missing value:
Normal data distribution → mean can be used
skewed data distribution → median.

5) use the attribute mean or median for all samples belonging to same class as the given tuple.
ex: credit card risk classification.

6) Use the most probable value to fill in the missing value:
→ determind by with regression, decision tree, Bayesian formulism.

◇ Noisy data

Noise - random error or variance in a measured variable

ways
- faulty data collection instruments
- data entry problems.
- data transmission problems.
- technology limitation.
- inconsistency in naming convention.

Methods to handle noisy data:

- Binning
1) sort data
2) partition into equal frequency bins.
3) a) smooth by bin means, each value in bin is replaced by mean value of bin.
b) smooth by bin medians: replaced by median.
c) smooth by bin boundries: replaced by closest boundry value.

Page no.: ____/
Date: ____/____/____/

- Regression.
  Smooth by fitting data into regression function.
  - linear regression
  - multiple linear regression

- Outlier analysis: Detected by "Clustering"

- Combined computer & human inspection.
  Detect suspicious values & check by human.

Data Integration:

→ merging of data from multiple data stores.
→ Careful integration can help reduce & avoid redundancies & inconsistancies.
→ improve speed & accuracy of DM process.

(I) Entity identification problem:

- data analysis involves data integration which combines data from multiple sources into data warehouse.

{ issues for data integration → schema integration
                              → object matching
↳ These are called as Entity Identification problem

when matching attributes from one db to another db, special attention must be paid to "structure" of data.
→ This ensures any attributes func. dependencies & constraints match to target system.

ex: discount in each order → discount in each item
SOURCE ——→ TARGET SYS

Page no.:
Date:

② Redundancy and Correlation Analysis.

Redundant data can occur when integration

- Object identification : Same attribute, different names.
- Derivable data : One attribute may be "derived" attribute in another table.

Solution:

Some redundancies can be detected by Correlation Analysis ~~or~~ AND covariance Analysis.

↓

Such analysis can measure how strongly one attribute implies the other.

For nominal data → χ^2 test

numeric attributes → correlation coefficient and covariance.

③ Tuple duplication

→ Duplication should also be detected at tuple level.

→ Use of denormalized tables is another source of data redundancy.

→ Inconsistencies arise between duplicates

ex: purchase order db contains attributes for the purchaser's name and address instead of key to this information.

↳ descrepancies can occour.

(4) Data value Conflict Detection and Resolution:

→

* ~~ex:~~ Data Attribute values from different sources may differ.
This maybe due to diffences in representation, scaling or encoding.

ex: One university may adopt a quater system, offer 3 courses on DBMS and assign grades A to F
while other
University may opt semester system, 2 courses DBMS, assign grads 1 to 10.

It is difficult to work out precise course to grade transformation rules betwn 2 universities.

* → Attributes can also differ ~~in~~ at abstraction level.
ex: totalsales in one db → refer to one branch of company
while other db
total_sales refer to all branch.

Data Reduction

→ Techniques can be applied to obtain a reduced representation of the data set that is much smaller, yet closely maintains the integrity of og data.
∴ mining on reduced data should be more effecient yet produce the same results.

M ♡ D

Page no.:
Date: / /

Data Reduction Strategies :

1. → Dimensionality reduction
2. → Numerosity reduction
3. → Data compression

(1) Process of reducing the number of random variables or attributes under consideration.
- wavelet transforms
- principal component analysis } Transform og data onto a smaller space.

- Attribute subset selection -
  ↳ irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

(2) Replaces the og data volume by alternative, smaller forms of data representation.
ex: regression → ↳ parametric - only data params stored
ex: Histograms, → ↳ non parametric -
clustering, sampling, data cube ~~aggregation~~ aggregation.

(3) Transformations are applied so as to obtain a reduced or "compressed" representation of og data.
→ if og data can be reconstructed from the compressed data w/o any info loss
⇓
lossless.

→ we can reconstruct only approx. of og data ⇒ lossy.

Page no.:
Date: / /

Attribute Subset Selection (Refer ref book)

→ Data sets may contain hundreds of attributes, many are irrelevant to mining task or redundant. Attribute Subset Selection removes it.

↓

goal → Find min set of attributes such that resulting probability distribution of the data classes is as close as possible to og distribution.

⇒ greedy heuristic methods that explore a reduced search space are commonly used for Attri Subset Selec..

methods:

1. Stepwise forward selection:
   ① empty set
   ② Best of og attributes is determined and added to reduced set.
2. Stepwise backward elimination
   ① Full set of attributes
   ② removes worst attributes.
3. Combination of forward & backward elimination.
   ① At each step procedure selects the best attribute & removes worst.
4. Decision tree induction.
   ID3, C4.5, CART originally intended for classification,
   → flowchart like
   - internal node - denotes test on an attribute, branch denotes outcome.
   - leaf node - class prediction.

Histograms

→ Use binning ★

If each bucket represents only a single attribute-value/frequency pair, buckets are called singleton buckets.

Partition Rules:
Equal width → equal bucket range
Equal-freq → (or equal depth)

Adv: effective at approximating both dense & sparse data, highly skewed & uniformed data.

Clustering (Refer PPT)

→ consider data tuples as Obj-
① partition objects into groups / clusters based on
② (similarity) ~~is defined~~.
③ (quality of) a cluster by diameter.
(centroid)
④

→ can be effective if data is clustered but not if data is "smeared".

→ Can have hierarchical clustering & be stored in multidimensional index tree structures.

om the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View. Enable Editing

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:
- Note: Sampling may not reduce database I/Os (page at a time)

Notes Comments

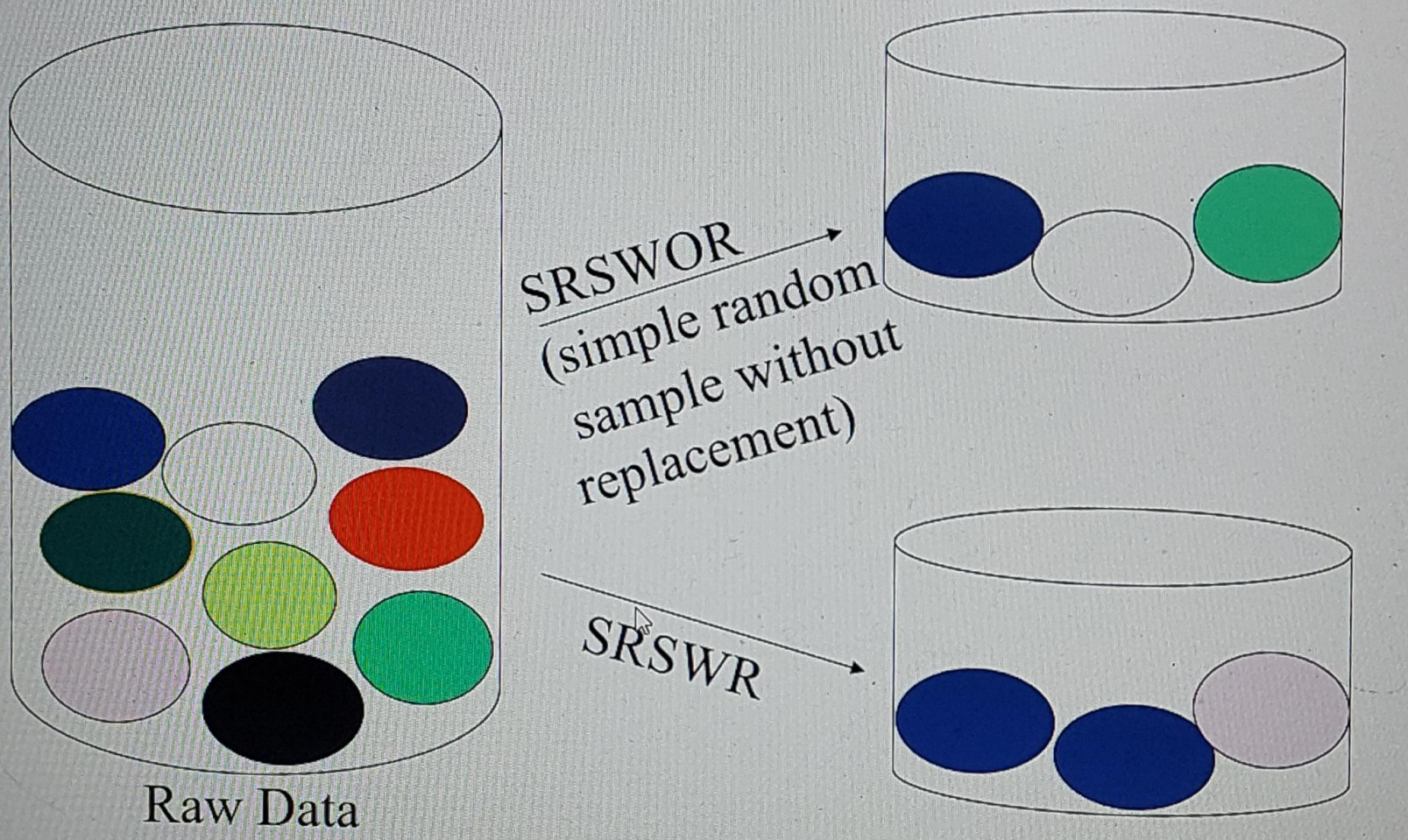
Unless you need to edit, it's safer to stay in Protected View. Enable Editing

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

Unless you need to edit, it's safer to stay in Protected View. Enable Editing

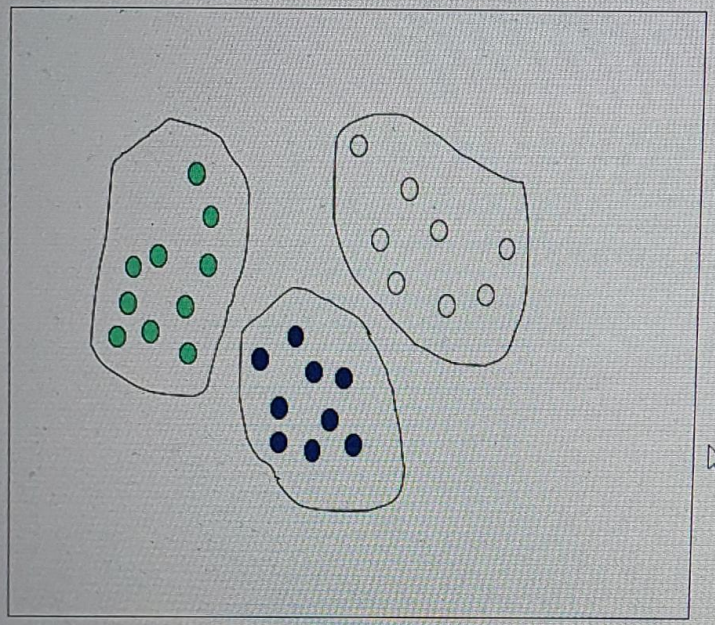
Sampling: With or without Replacement



Unless you need to edit, it's safer to stay in Protected View. Enable Editing

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample