Data Preparation & Cleaning

Anna Monreale
Computer Science Department

Introduction to Data Mining, 2nd Edition

Data understanding vs Data preparation

Data understanding provides general information about the data like

- The existence of **missing values**
- The existence of **outliers**
- the character of attributes
- **dependencies** between attributes.

Data preparation uses this information to

- select attributes,
- reduce the data dimension,
- select records,
- treat missing values,
- treat outliers,
- integrate, unify and transform data
- improve data quality

Data Preparation

- Aggregation
- Data Reduction: Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - **Data reduction**
    - Reduce the number of attributes or objects
  - **Change of scale**
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - **More "stable" data**
    - Aggregated data tends to have less variability

Example: Precipitation in Australia

- This example is based on precipitation in Australia from the period 1982 to 1993.

  The next slide shows

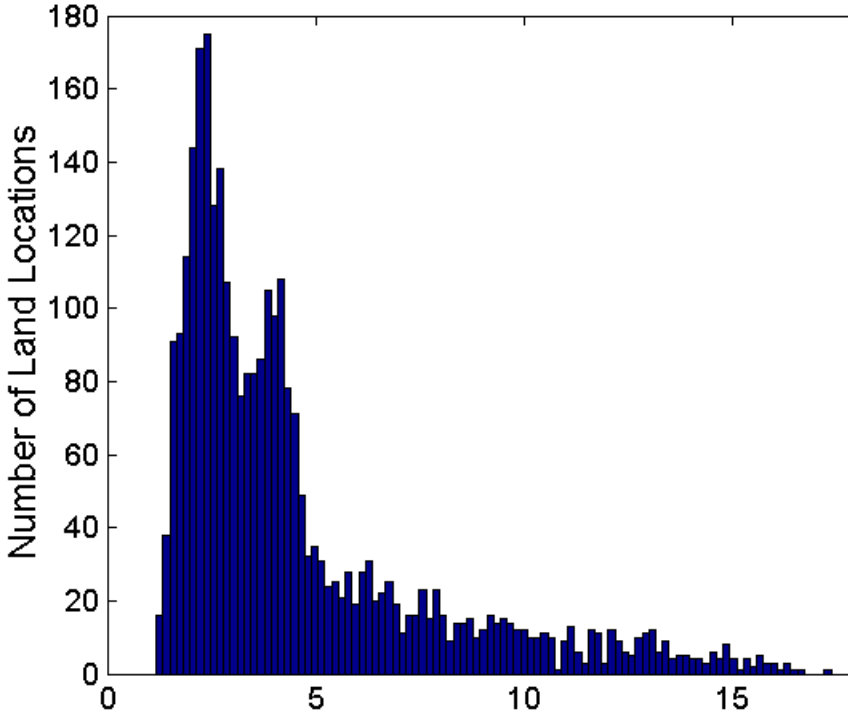
  – A histogram for the standard deviation of average monthly precipitation for specific locations in Australia, and

  – A histogram for the standard deviation of the average yearly precipitation for the same locations.

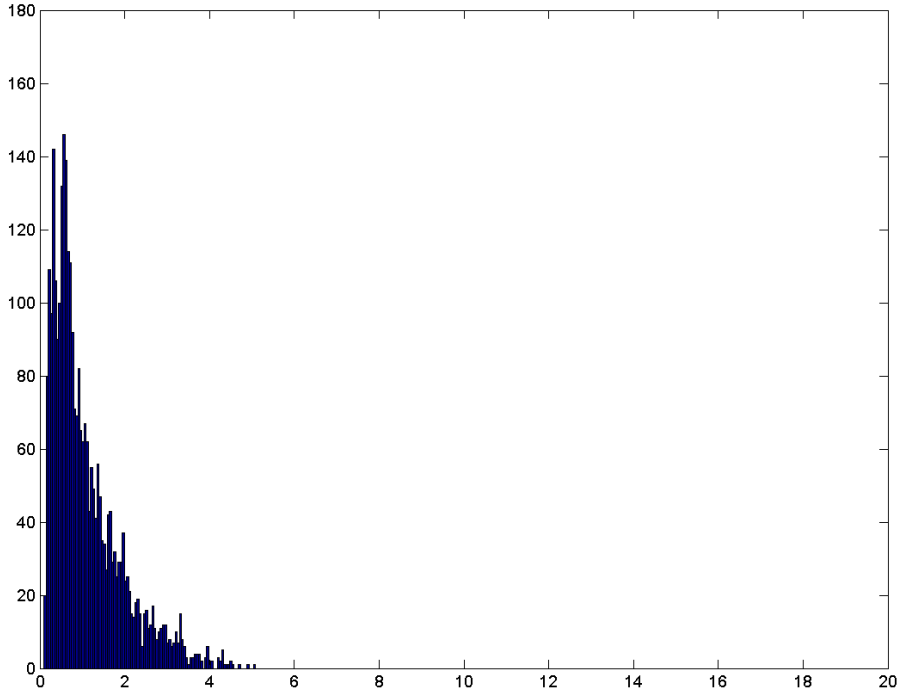
- The average **yearly precipitation has less variability** than the **average monthly precipitation.**
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia …

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Data Reduction

Reducing the amount of data

- Reduce the number of **records**
  - Data Sampling
  - Clustering
- Reduce the number of **columns** (attributes)
  - Select a subset of attributes
  - Generate a new (a smaller) set of attributes

Sampling

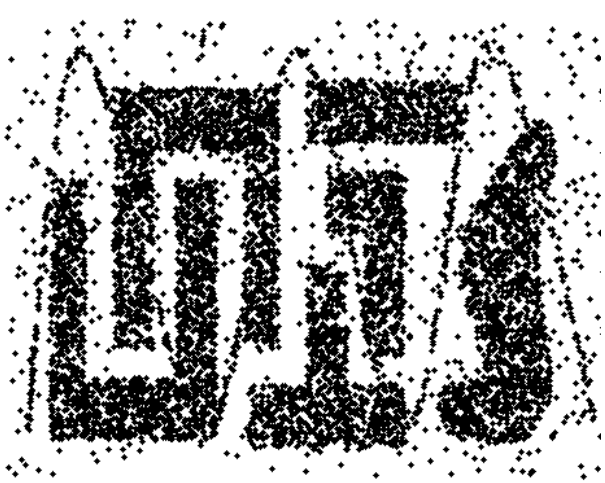
- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

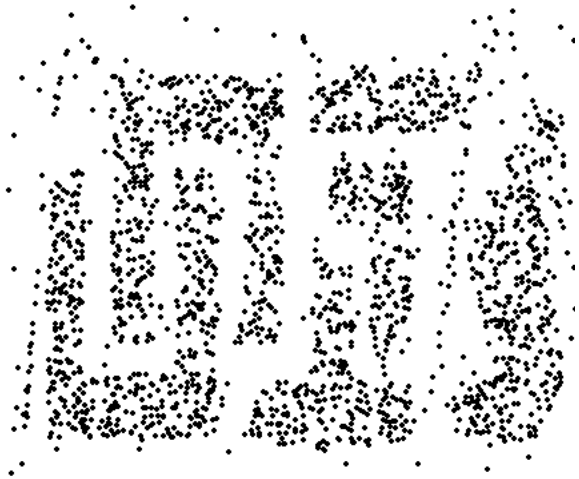
Sampling …

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set**, if the sample is representative**
  - A sample is representative if it has approximately the **same properties** (of interest) as the original set of data

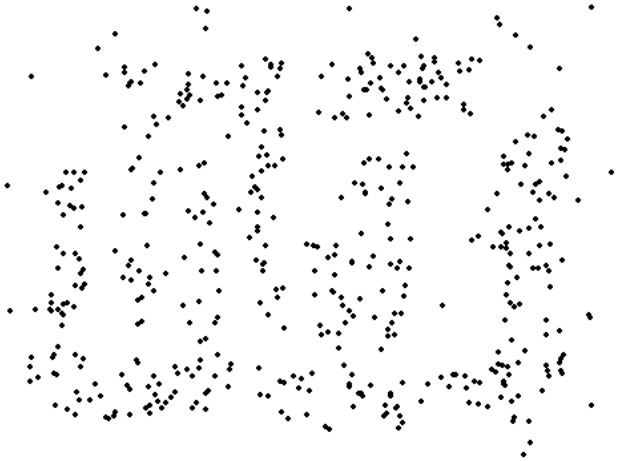
Sample Size



8000 points



2000 Points



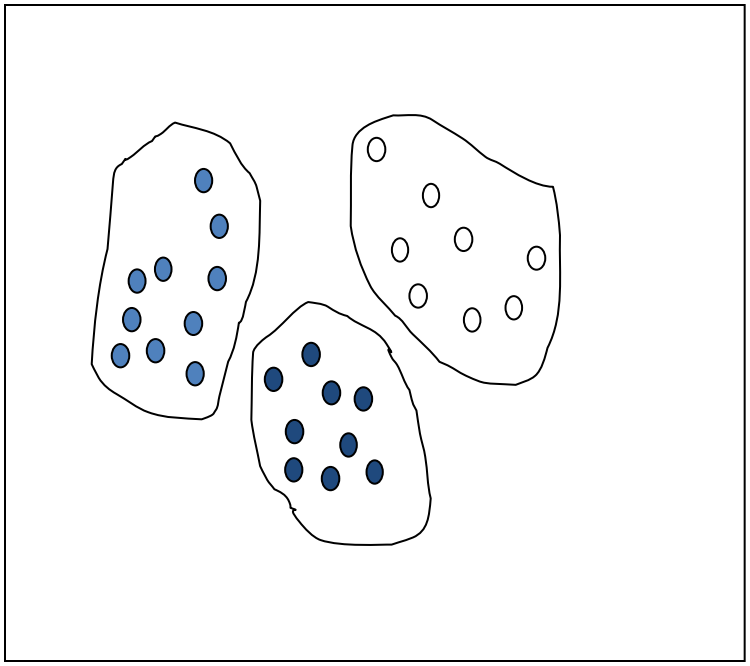
500 Points

Types of Sampling

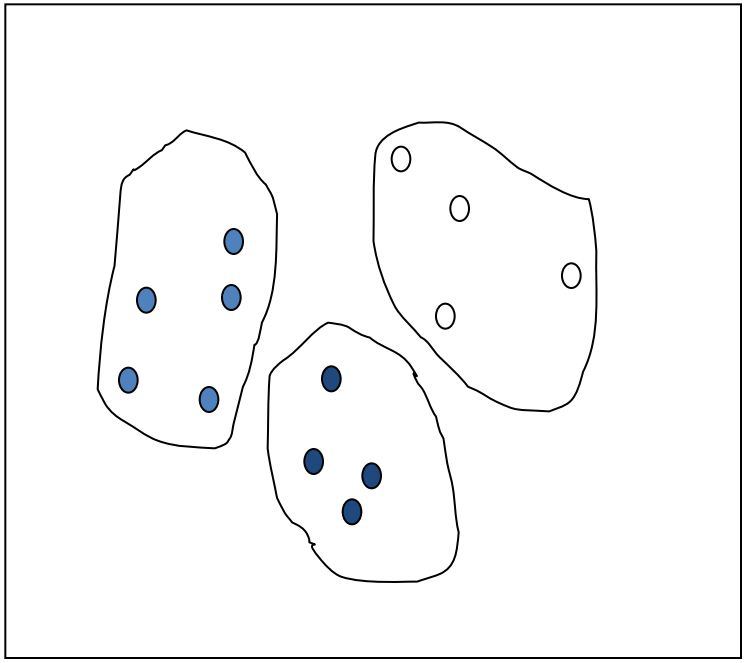
- Simple Random Sampling
  - There is an **equal probability** of selecting any particular item
  - **Sampling without replacement**
    - As each item is selected, it is removed from the population
  - **Sampling with replacement**
    - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition
  - Approximation of the percentage of each class
  - Suitable for distribution with peaks: each peak is a **layer**

Stratified Sampling

Raw Data



Cluster/Stratified Sample



Reduction of Dimensionality

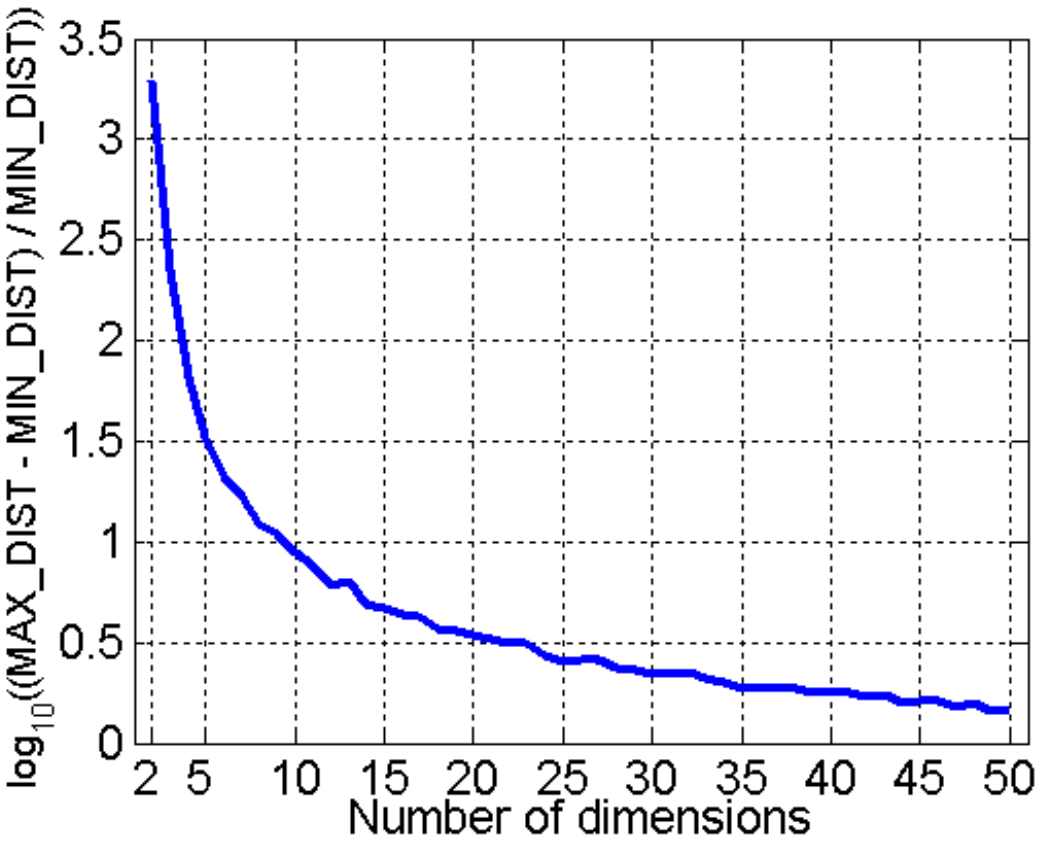
Selection of a subset of attributes that is as small as possible and sufficient for the data analysis.

- removing (more or less) **irrelevant** features
  - Contain **no information** that is **useful** for the data mining task at hand
  - **Example**: students' ID is often irrelevant to the task of predicting students' GPA
- removing **redundant** features
  - **Duplicate** much or all of the **information** contained in one or more other attributes
  - **Example**: purchase price of a product and the amount of sales tax paid

Curse of Dimensionality

- When dimensionality increases, data becomes **increasingly sparse** in the space that it occupies

- **Definitions of density and distance** between points, which are critical for clustering and outlier detection, **become less meaningful**



Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Removing irrelevant/redundant features

- For **removing irrelevant features**, a **performance measure** is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task

- For removing **redundant features**, either a performance measure for subsets of features or a correlation measure is needed.

Reduction of Dimensionality

Filter Methods

- Selection after analyzing the **significance** and **correlation** with other attributes
- Selection is independent of any data mining task
- The operation is a pre-processing

Wrapper Methods

- Selecting the top-ranked features using as reference a DM task
- Incremental Selection of the “best” attributes
- “Best” = with respect to a specific measure of statistical significance (e.g.: information gain)

Embedded Methods

- Selection as part of the data mining algorithm
- During the operation of the DM algorithm, the algorithm itself decides which attributes to use and which to ignore (e.g. Decision tree)

Wrapper Feature Selection Techniques

- **Selecting the top-ranked features**: Choose the features with the best evaluation when single features are evaluated.

- **Selecting the top-ranked subset:** Choose the subset of features with the best performance. This requires exhaustive search and is impossible for larger numbers of features. (For 20 features there are already more than one million possible subsets.)

- **Forward selection:** Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.

- **Backward elimination:** Start with the full set of features and remove features one by one. In each step, remove the feature that yields to the least decrease in performance.

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature construction
    - Domain-dependent
    - Example: dividing mass by volume to get density
  - Feature Projection
    - transforms the data from the high-dimensional space to a space of fewer dimensions

Feature Creation: features needed for task

Find the best workers in a company.

- Attributes :
  - the tasks, a worker has finished within each month,
  - the number of hours he has worked each month,
  - the number of hours that are normally needed to finish each task.
- These attributes *contain* information about the efficiency of the worker.
- But instead using these three "raw" attributes, it might be more useful to define a new attribute *efficiency*.
- efficiency $= \frac{\text{hours actually spent to finish the tasks}}{\text{hours normally needed to finish the tasks}}$

Feature Creation: features needed for task

- **Task**: face recognition in images
- Images are only set of contiguous pixels
- They are not suitable for many types of classification algorithms
- Process to provide **higher level features**
  - presence or absence of certain types of areas that are highly correlated with the presence of human faces
  - a much broader set of classification techniques can be applied to this problem

Feature Projection or Extraction

- It transforms the data in the high-dimensional space to a space of fewer dimensions.
- The data transformation may be linear, or nonlinear.
- Approaches:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Non-negative matrix factorization (NMF)
  - Linear Discriminant Analysis (LDA)
  - Autoencoder