Data Mining:

- Introduction
- Data
- Data Preprocessing

OUTLINE

- Introduction
  1. Why Data Mining ?
  2. What Is Data Mining ?
- Data
  1. Data Object and Attribute Types
  2. Data Quality
  3. Measuring Data Similarity and Dissimilarity
  4. Data Visualization
- Data Preprocessing
  1. Aggregation
  2. Sampling
  3. Dimensionality Reduction
  4. Feature subset selection
  5. Feature creation
  6. Discretization and Binarization
  7. Attribute Transformation

WHY DATA MINING ?

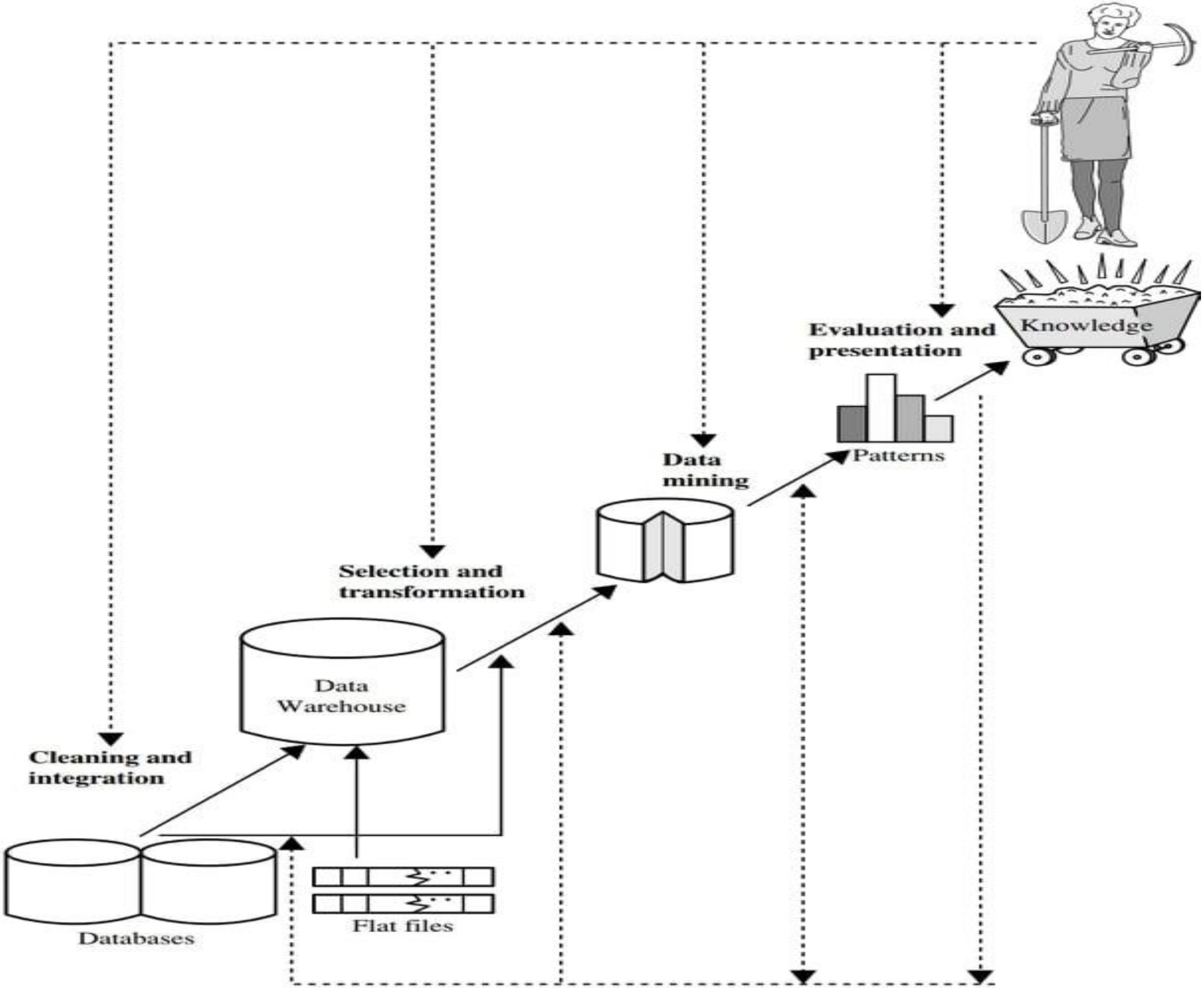
- We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need.
  - Major sources of abundant data
    - Business
    - Science
    - Society
    - Engineering
    - Medicine
- Data rich but information poor!

What Is Data Mining ?

- Extracting knowledge from large amounts of data.

knowledge discovery from data (KDD):

- **Data cleaning** (remove noise and inconsistent data)
- **Data integration** (multiple data sources may be combined)
- **Data selection** (data relevant to the analysis task are retrieved from database)
- **Data transformation** (data transformed and consolidated into forms appropriate for mining)
- **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **Pattern evaluation** (indentify the truly interesting patterns)
- **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)


Knowledge
Evaluation and presentation
Patterns
Data mining
Selection and transformation
Data Warehouse
Cleaning and integration
Databases
Flat files

Data Object and Attribute Types

- **Data** : Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Attribute is also known as variable, field, characteristic, or feature
  - Object is also known as record, point
  - Database rows -> data objects; columns ->attributes.
- Type of attribute
  - **Nominal**
  - **Binary**
  - **Ordinal**
  - **Interval**
  - **Ratio**

- **Nominal:** categories, states, or “names of things”
  - *nominal attributes provide only enough information to distinguish one object from another (=,≠)*
  - *Example : Hair_color,* 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**
  - The values of an ordinal attribute provide enough information to order objects (=, ≠, <, >)
  - *Size* = {*small, medium, large*}, grades

- **Interval**
  - For interval attributes, the differences between values are meaningful (=, ≠, <, >, +, -)
  - E.g., *temperature in C°or F°*
  - No true zero-point
- **Ratio**
  - For ratio variables, both differences and ratios are meaningful (=, ≠, <, >, +, - , *, /)
  - Inherent **zero-point**
  - E.g., *temperature in Kelvin, length, counts*

Discrete Attribute	Continuous Attribute
Has only a finite or countably infinite set of values	Has real numbers as attribute values
zip codes, counts	temperature, height, or weight
Sometimes, represented as integer variables	Practically, real values can only be measured and represented using a finite number of digits
Note: Binary attributes are a special case of discrete attributes	Continuous attributes are typically represented as floating-point variables

- Type of Data Sets:
  - Record: Data matrix, Document data and Transaction data
  - Graph
  - Ordered

Data Quality

- Examples of data quality problems:
  - Noise
    - Noise refers to modification of original values
  - Outliers
    - Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - Missing values
    - Reasons for missing values
      - Information is not collected (e.g., people decline to give their age and weight)
      - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
    - Handling missing values
      - Eliminate Data Objects
      - Estimate Missing Values
      - Ignore the Missing Value During Analysis
      - Replace with all possible values (weighted by their probabilities)
  - Duplicate data
    - Data set may include data objects that are duplicates
      - Ex: Same person with multiple email addresses
    - Data cleaning: Process of dealing with duplicate data

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale

Measuring Data Similarity and Dissimilarity

- Similarity
  - Numerical measure of the degree to which two data objects are alike.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of the degree to which two data objects are different.
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

- Proximity measures, especially similarities, are defined to have values in the interval [0,1].
- If the similarity between objects can range [min_s,max_s] .We can make them fall Into the range [0,1] by using the formula: $s' = \frac{s - \min(s)}{\max(s) - \min(s)}$
- Dissimilarity measures with a finite range [min_d, max_d] can be mapped to the interval [0,1] by using the formula: $d' = \frac{d - \min(d)}{\max(d) - \min(d)}$
- If the dissimilarity measure originally takes values in the interval [0, ∞], then we usually use the formula: $d' = \frac{d}{d+1}$ for such cases and bring the dissimilarity measure between [0,1].
- If the dissimilarity (or similarity) falls in the interval [0,1], then the similarity (or dissimilarity) can be defined as : s = 1- d (d = 1 - s).
- If dissimilarity fall in other ranges then the similarity can be defined as: $s = \frac{1}{d+1}$, $s = e^{-d}$,or $s = 1 - \frac{d - \min(d)}{\max(d) - \min(d)}$

	Similarity	Dissimilarity
Nominal	S=$\begin{cases} 1\ if\ x = y \\ 0\ if\ x \neq y \end{cases}$	d=$\begin{cases} 0\ if\ x = y \\ 1\ if\ x \neq y \end{cases}$
Ordinal	S=1-d	d=$\frac{\lvert x-y \rvert}{n-1}$

Interval or Ratio	S=-d, s=$\frac{1}{d+1}$, s=e^{-d},
	s=1-$\frac{\mathrm{d}-\min(\mathrm{d})}{\max(\mathrm{d})-\min(\mathrm{d})}$
		d=$\lvert x - y \rvert$

- Properties of Euclidean Distance
  - d(p, q) ≥0 for all p and q and d(p, q) = 0 only if p = q.
  - d(p, q) = d(q, p) for all p and q. (Symmetry)
  - d(p, r)≤ d(p, q) + d(q, r) for all points p, q, and r. (Triangle Inequality) where d(p, q) is the distance (dissimilarity) between points (data objects), p and q.

Measures that satisfy all three properties are known as metrics.

In Similarities Triangle Inequality not hold :

1) s(p,q)=1 if p=q

2) s(p,q) = s(q,p) for all p and q. (Symmetry)

- Similarity Between Binary Vectors:
  - Simple Matching Coefficient (SMC)=$\frac{f_{11}+f_{00}}{f_{00}+f_{11}+f_{01}+f_{10}}$
  - Jaccard Coefficient(J)=$\frac{f_{11}}{f_{11}+f_{01}+f_{10}}$

Example: x=(1,0,0,0,0,0,0,0,0,0) ,y=(0,0,0,0,0,0,1,0,0,1)

Solution: f_{11}=0 ; f_{00} = 7 ; f_{01} = 2 ; f_{10}=1

SMC=$\frac{0+7}{7+0+2+1}$; $J = \frac{0}{0+2+1}$

- Cosine Similarity: $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$

Example: Find the **similarity** between documents 1 and 2.

$d_1 =$ (5, 0, 3, 0, 2, 0, 0, 2, 0, 0),$d_2 =$ (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2$= 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25

$||d_1||$= $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}$= 6.481

$||d_2||$= $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}$ = 4.12

$\cos(d_1, d_2)$ = 0.94

General Approach for Combining Similarities:

For the 1)k^{th} $attribute, compute\ a\ similarity\ s_k\ in\ range[0,1]$

2)Define an indicator variable δ_k $for\ the\ k_{th}$ attribute as follows:

$$\delta_k = \begin{cases} 0\ if\ the k^{th}\ attribute\ is\ binary\ asymmetric \\ and\ both\ object\ have\ a\ value\ of\ 0\ , \\ or\ if\ one\ of\ the\ objects\ has\ a\ missing\ values\ for\ the k^{th}\ attribute \\ 1\ otherwise \end{cases}$$

3) similarity(p,q)=$\frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$

Data Visualiztion

- Visualization is the conversion of data into a visual or tabular format.
  - Visualization of data is one of the most powerful and appealing techniques for data exploration.
- Representation: Mapping Data to Graphical Elements:
  - Is the mapping of information to a visual format
  - Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.

Ex:

- Objects are often represented as points
- Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape

- Arrangement : Is the placement of visual elements within a display
- Selection : Is the elimination or the de-emphasis of certain objects and attributes
  - Selection may involve the choosing a subset of attributes
  - Selection may also involve choosing a subset of objects
- Visualizing Small Numbers of Attributes:
  - the distribution of the observed values for a single attribute such as histograms
  - the relationships between the values of two attributes such as scatter plots
- Stem and Leaf Plots: we split the values into groups, where each group contains those values that are the same except for the last digit. Each group becomes a stem, while the last digits of a group are the leaves.
  - Ex: 43 44 44 45 46 51 52 53 53 62 61
    - 4 : 34456
    - 5 : 1233
    - 6 : 21

- Histograms :
  - the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - Shape of histogram depends on the number of bins
- Box plots :
  - Box plots are another method for showing the distribution of the values of a single numerical attribute.
    - The lower and upper ends of the box indicate the 25 and 75 percentiles
    - The line inside the box indicates the value of the 50 percentile.
    - The top and bottom lines of the tails indicate the 10 and 90 percentiles.
    - Outliers are shown by "+" marks.
- Pie Chart :
  - A pie chart is similar to a histogram, but is typically used with categorical attributes
  - a pie chart uses the relative area of a circle to indicate relative frequency.

- Scatter Plots:
  - Most commom in Scatter Plot Two-dimensional
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- Extending Two and Three-Dimensional Plots :
  - Scatter plots can display up to three additional attributes using color or shading, size, and shape, allowing five or six dimensions to be represented.
  - As the complexity of a visual representation of the data increases, it becomes harder for the audience to interpret the information.
  - There is no benefit in packing six dimensions' worth of information into a two or three-dimensional plot, if doing so makes it impossible to understand
- Contour plots :
  - Useful when a continuous attribute is measured on a spatial grid
  - They partition the plane into regions of similar values
- Vector Field Plots
  - In some data, a characteristic may have both a magnitude and a direction associated with it.

- Parallel Coordinates :
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - each object is represented as a line
  - Parallel coordinates have one coordinate axis for each attribute, but the different axes are parallel to one other instead of perpendicular
- Star Plots :
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon
- Chernoff Faces
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face

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
- More “stable” data
  - Aggregated data tends to have less variability

Sampling

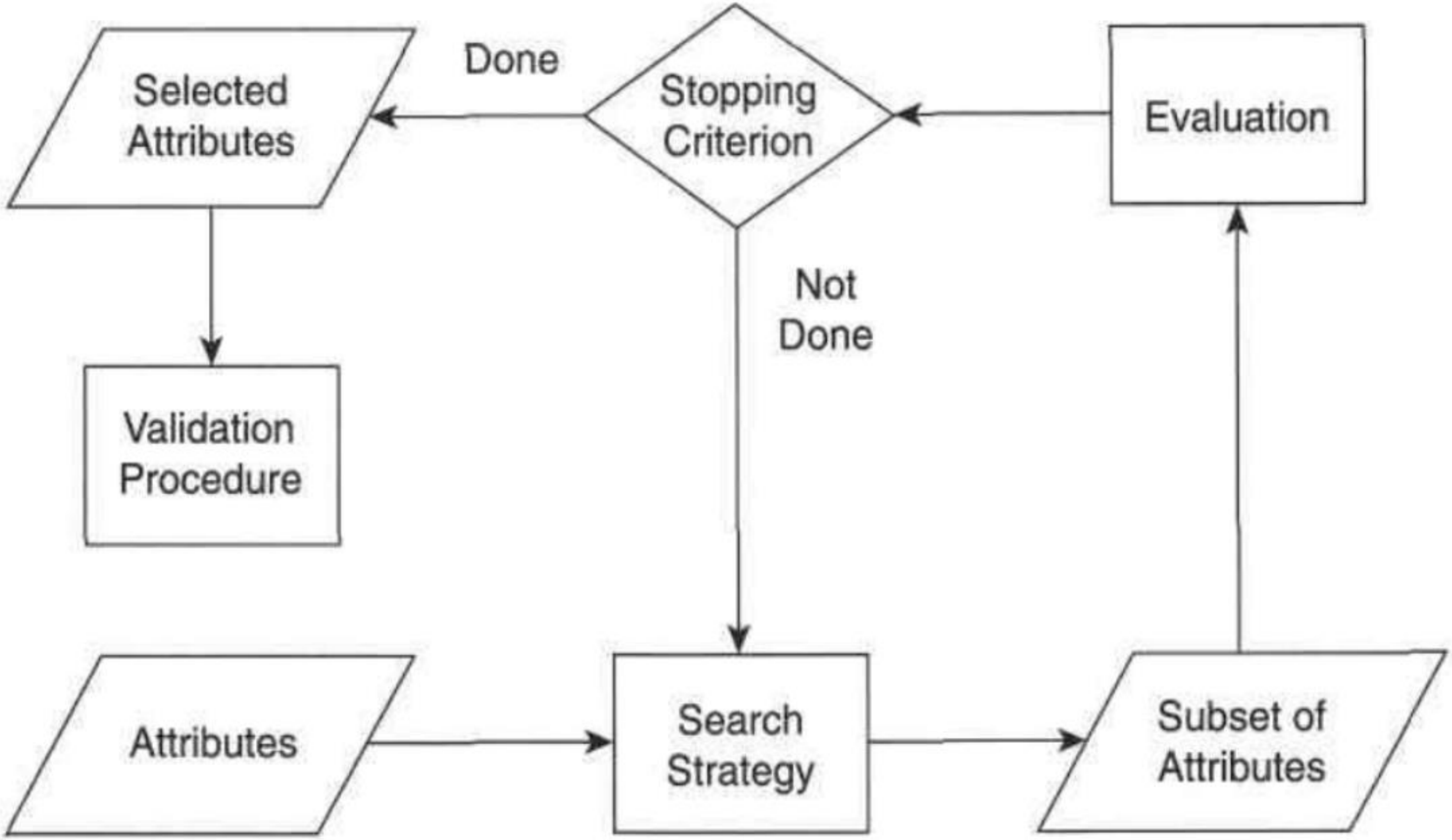
- Sampling is the main technique employed for data selection
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.
- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets
- 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.
  - Stratified sampling
    - Split the data into several partitions; then draw random samples from each partition

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality(When dimensionality increases, data becomes increasingly sparse in the space that it occupies)
  - 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
  - Principle Component Analysis(PCA)
  - Singular Value Decomposition

Feature Subset Selection

- Another way to reduce dimensionality of data
- 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
- 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
- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes


Selected Attributes
Done
Stopping Criterion
Evaluation
Not Done
Validation Procedure
Attributes
Search Strategy
Subset of Attributes

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 Extraction
  - Mapping Data to New Space
  - Feature Construction
- features constructed out of the original features can be more useful than the original features.
  - Ex: density=$^{mass}/_{volume}$

Discretization and Binarization

- Discretization : to transform a continuous attribute into a categorical attribute.
  - Divided continuous attribute into n interval by n-1 split points
  - All the values in one interval are mapped to the same categorical value
- Binarizaion: to transformed continuous and discrete attributes into binary attributes.
  - If there are n categorical values ,each original value to an integer in interval [0,n-1],Next convert each of these n integer to a binary number . Ex: categorical values are {awful , poor ok ,good ,great}

categorical	Integer	x_1	x_2	x_3
awful	0	0	0	0
Poor	1	0	0	1
Ok	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k , log(x), e^x, |x|

REFERENCES :

1. PANG-NING Tan, MICHAEL STEINBACH , and VIPIN KUMAR , INTRODUCTION TO DATA MINING

2. Jiawei Han ,Micheline, and Jian Pei , DATA MINING Concepts and Techniques,Third Edition