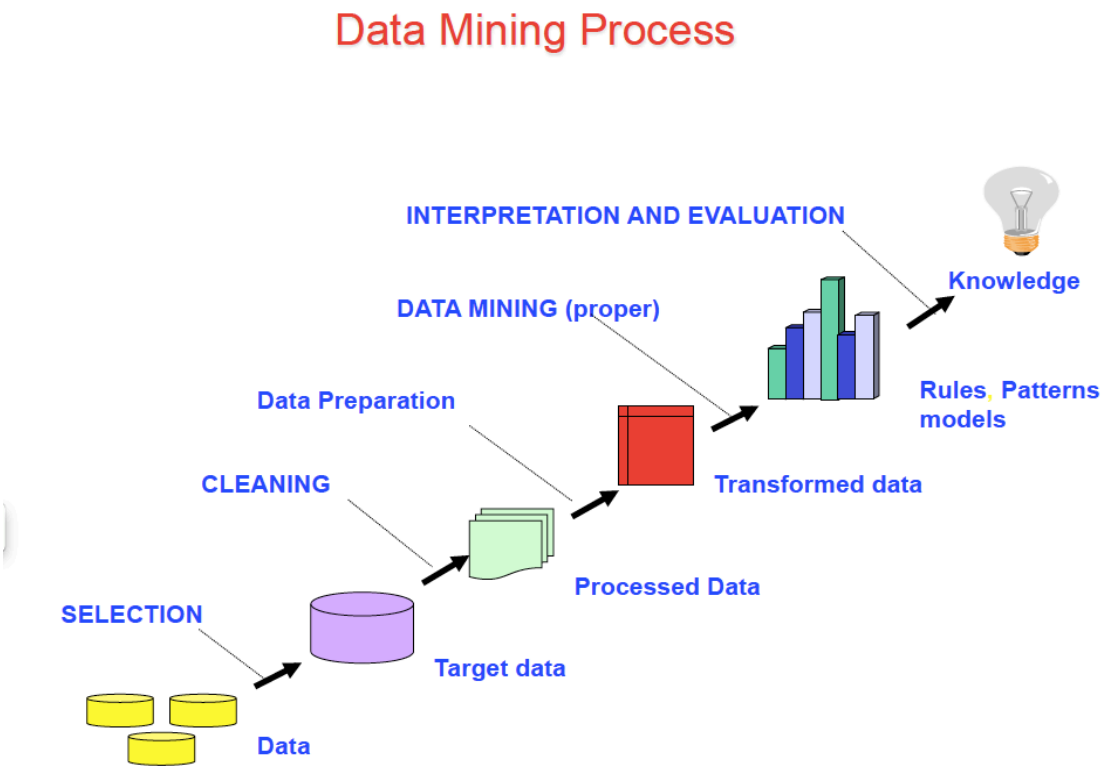
1. Data Mining and Learning Process



Preprocessing and DM proper see below.
Interpretation: discovered patterns are presented in a proper format and the user decides if it is necessary to re-iterate the algorithms.

2. Preprocessing stage

Preprocessing: includes all the operations that have to be performed before a data mining algorithm is applied
Data in the real world is dirty: incomplete, noisy and inconsistent. Quality decisions must be based on quality Data.
Data cleaning
– Fill in missing values, smooth noisy data(binning, clustering, regression), identify or remove outliers, and resolve inconsistencies
Data integration
– Integration of multiple databases, data cubes, or files
Data transformation
Normalization and aggregation
Data reduction and attribute selection
Obtains reduced presentation in volume but produces the same or similar analytical results (stratified sampling, PCA, cluster)
Data discretization
Part of data reduction but reduces the number of values of the attributes by dividing the range of attributes into intervals (segmentation by natural partition, hierarchy generation)

3. Data mining proper

DM proper is a step in the DM process in which algorithms are applied to obtain patterns in data.
It can be re-iterated- and usually is.

4. Descriptive/ non descriptive data mining and models

Statistical - descriptive.

- Statistical data mining uses historical data to predict some unknown or missing numerical values.
- Descriptive data mining aims to find patterns in the data that provide some information about what the data contains.
  often presents the knowledge as a set of rules of the form IF.... THEN...
  In this case it is called a Descriptive DM

We often define the concept – CLASS by distinguishing in our database an attribute C and its value v. In this case the concept – class description is written
C = {records: C=v}
We say that C is a **CLASS** with the description C=v
We call C a class attribute. Let C has a values v1, v2, … vk. In this case C define k classes C1, C2, … Ck:

C1 = {records: C=v1}, C2 = {records: C=v1}, …. C2 = {records: C=vk}

Discriptive: Decision Trees, Rough Sets, Classification by Association
Statistical: Neural Networks, Bayesian Networks, Cluster, Outlier analysis, Trend and evolution analysis
Optimization method: Genetic Algorithms – can be descriptive

5. What and how to decide which type of data mining to use

Different Data Mining methods are required for different kind of data and different kinds of goals

6. Application and algorithms for them

Business Advantages
Data Mining uses gathered data to predict tendencies and waves, to classify new data, to find previously unknown patterns for the use for business advantages, to discover unknown relationships

Fraud Detection and Management
use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
1.auto insurance: detect characteristics of group of people who stage accidents to collect on insurance
2.money laundering: detect characteristics of suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
3.medical insurance: detect characteristics of fraudulent patients and doctors
Detecting inappropriate medical treatment
4.Detecting telephone fraud
– DM builds telephone call model: destination of the call, duration, time of day or week.
– Detects patterns that deviate from an expected norm.

Market Analysis and Management
Target marketing
– DM finds clusters of "model" customers who share the same characteristics: interest,

income level, spending habits, etc. Determine customer purchasing patterns over time
Customer profiling
– data mining can tell you what types of customers buy what products (clustering or classification)

According to Algorithms:
Classification:
– classify countries based on climate
– classify cars based on gas mileage and use it to predict classification of a new car
Cluster analysis
cluster houses to find distribution patterns
Outlier analysis
It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

7. Classification

Classification:
Finding models (rules) that describe (characterize) or/ and distinguish (discriminate) classes or concepts for future prediction
Classification Data Format:
a data table with key attribute removed.
Special attribute, called a class attribute must be distinguished.
The values: c1, c2, ...cn of the class atrribute C are called class labels. The class label attributes are discrete valued and unordered.
Goal:
FIND a minimal set of characteristic and/or discriminant rules, or other descriptions of the class C, or (all) other classes.
We also want the found rules to involve as few attributes as it is possible

8. Classification process

Stage 1: build the basic patterns structure-**training**
Stage 2: optimize parameter settings; can use (N:N) re-substitution- parameter tuning
Re-substitution error rate = training data error rate
Stage 3: use **test data** to compute- predictive accuracy/error rate - **testing**

9. Classification models and differences

Decision Trees –descriptive
Discovering discriminant rules
Method: successive division of the set of data

Neural Networks- statistical
the network is trained to obtain classification patterns

Bayesian Networks - statistical
is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).

Rough Sets - descriptive
is a formal approximation of a conventional set in terms of a pair of sets which give the *lower* and the *upper* approximation of the original set

Genetic Algorithms – descriptive or statistical
Mimics the process of natural selection.is routinely used to generate useful solutions to optimization

10. Decision tree induction

A flow-chart-like tree structure ;
Internal node denotes an attribute;
Branch represents the values of the node attribute;
Leaf nodes represent class labels

The **basic DT algorithm** for decision tree construction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner.
Tree STARTS as a single node representing all training dataset (data table with records called samples)
IF the samples (records in the data table) are all in the same class, THEN the node becomes a leaf and is labeled with that class
The algorithm uses the same process recursively to form a decision tree at each partition
The recursive partitioning STOPS only when any one of the following conditions is TRUE
1. All records (samples) for the given node belong to the same class
2. There are no remaining attributes on which the samples (records in the data table) may be further partitioned
Majority voting involves converging node N into a leaf and labeling it with the most common class in **D** which is a set of training tuples and their associated class labels
3. There is no records (samples) left – a LEAF is created with majority vote for training sample

Heuristics: Attribute Selection Measures
Information gain, Gini index
For selecting the attribute that "best" discriminates the given tuples according to class

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$$

GainRatio(A)=Gain(A)/SplitInfo(A)
Maximum gain ratio is selected as the splitting attribute

An induced tree may overfit the training data: pre/post pruning

Why decision tree induction in data mining?
– relatively faster learning speed (than other classification methods)
– simple and easy to understand – descriptive - classification rules
– can use SQL queries for accessing databases

– comparable classification accuracy with other methods

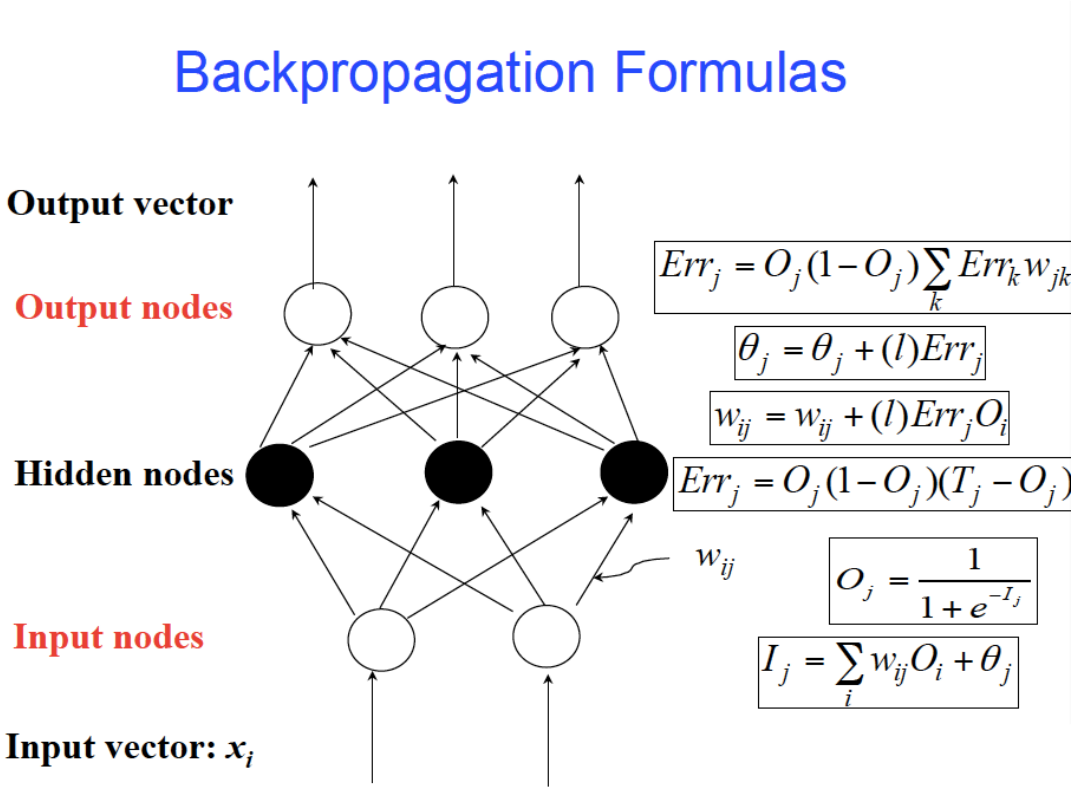
11. Neutral network model, strength and weakness

Neural Network is a set of connected INPUT/OUTPUT UNITS, where each connection has a WEIGHT associated with it. Neural Network learns by adjusting the weights so as to be able to correctly classify the training data and hence, after testing phase, to classify unknown data.
Neural Network needs long time for training. Determining network topology is difficult. Choosing single learning rate impossible.(train with subset)

Neural Network has a high tolerance to noisy and incomplete data. generally better with larger number of hidden units

The inputs to the network correspond to the attributes and their values for each training tuple
Inputs are fed simultaneously into the units making up the input layer
Inputs are then weighted and fed simultaneously to a hidden layer
The number of hidden layers is arbitrary, although often only one or two
The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction.

For each training sample, the weights are first set random then they are modified as to minimize the mean squared error between the network's classification (prediction) and actual classification.
Backpropagation Algorithm:
STEP ONE: initialize the weights and biases
STEP TWO: feed the training sample
STEP THREE: propagate the inputs forward
STEP FOUR: backpropagate the error
STEP SIX: repeat and apply terminating Conditions



Stops when:
All wij in the previous epoch are below some threshold
The percentage of samples misclassified in the previous epoch is below some threshold
a pre- specified number of epochs has expired

12. Classifier

For the reason that we can use discovered patterns (discriminant and /or characteristic rules) to classify unknown sets of objects, a classification algorithm is often called **shortly** a classifier
Name Classifier implies more than just a classification algorithm.

A Classifier is a final product of a process that uses data set and a classification algorithm.

13. Building a classifier

Building a classifier consists of two phases: training and testing.
We use the training data set to create patterns: rules, trees, or to train a Neural or Bayesian network.
We evaluate created patterns with the use of test data.
We terminate the process if it has been trained and tested and the predictive accuracy is on an acceptable level.

PREDICTIVE ACCURACY of a classifier is a percentage of well classified data in the test data set.

Basic methods of training and testing:
The main methods of predictive accuracy evaluations are:
• Resubstitution (N ; N)

• Holdout (2N/3 ; N/3)
• k-fold cross-validation (N- N/k ; N/k)
• Leave-one-out (N-1 ; 1)

14. AssociationAnalysis:

Finding frequent patterns called associations, among sets of items or objects in transaction databases, relational databases, and other information repositories

Confidence:

The rule X->Y holds in the database D with confidence c if the c% of the transactions in D that contain X also contain Y

Support:

The rule X->Y has support s in D if s% of the transaction in D contain XUY

We (user) fix MIN support usually low and Confidence high

$$conf(A \Rightarrow B) = \frac{sc(A \cup B)}{scA}$$

$$Support(A \Rightarrow B) = P(A \cup B) = \frac{sc(A \cup B)}{\#D}$$