Cse537 Lecture Notes

Introduction to Learning

Professor Anita Wasilewska
Computer Science Department
Stony Brook University

Learning Main Objectives

- **Indentification** of data as a source of useful information, called also a knowledge

- **Use** of “learned” information (knowledge) for different applications

Data – Information - Knowledge

- **Data – as in databases**
- **Information, or knowledge is a meta information ABOUT the patterns hidden in the data**
- **The patterns must be discovered automatically**

Learning : Intuitive Definition

- **Learning** is a process that extracts previously unknown knowledge from the data

- It requires special algorithms, technologies and methods

Learning

- There are many types of learning.
- We will cover two:
- SUPERVISED LEARNING: classification
- UNSUPERVISED LERANING: clustering
- The knowledge obtained in the learning process is often presented as a set of rules of the form:

  IF.... THEN…..

In this case it is called Descriptive Learning

- It also finds other relationships in data

Some Commercial Applications

- **Market analysis and management**
    - target marketing, customer relation management
    - **Risk analysis and management**
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis

More Applications

- Buying patterns
- Fraud detection
- Customer Campaings
- Decision support
- Medical aplications
- Marketing
- and more

Fraud Detection and Management

- **Applications**

  widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.

- **Approach**

  use historical data to build models of fraudulent behavior and use learned knowledge to help identify similar instances

Fraud Detection and Management

- Examples (historical)

  auto insurance: learn characteristics of group of people who stage accidents to collect on insurance and use them automatically to prevent fraud

  money laundering: learn characteristics of suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)

  medical insurance: learn characteristics of fraudulent patients and doctors

Fraud Detection and Management

- **Detecting telephone fraud**

  Use learning methods to describe telephone call model: destination of the call, duration, time of day or week. Detects patterns that deviate from an expected norm.

  British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.

- **Detecting Credit Card fraud**

  Use learning methods to describe a given person (or general) credit card usage model.

  Detect patterns that deviate from an expected norm.

Market Analysis and Management

- **Customer profiling**

We use learning algorithms (clustering or classification) to identify:

1. what types of customers buy what products;
2. customer preferences;
3. the best products for different customers

Business Summary

- **Learning Process** (called also Data Mining in a case of very large data sets)
- helps to **improve competitive advantage** of organizations in dynamically changing environment;
- **it improves clients retention** and **conversion**
- **Different methods** are needed for different kind of **data** and different **goals**

Scientific Applications

- Networks failure detection
- Controllers
- Geographic Information Systems
- Genome- Bioinformatics
- Intelligent robots
- Intelligent rooms
- etc… etc ….

What is NOT Learning

- Once the **patterns** are FOUND and TESTED the **learning process is finished**
- Use of the patterns **is not** Learning
- Querries to the database **are not** Learning

Evolution of Database Technology

- 1960s:

  Data collection, database creation, IMS and network DBMS

- 1970s:

  Relational data model, relational DBMS implementation

Evolution of Database Technology c.d.

- 1980s:

  RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)

- 1990s—2000s:

  Data mining **- learning** is an integral part of it) and data warehousing, multimedia databases, and Web databases

  BIG DATA **- learning** is also an integral part of it – new name - NEW algorithms, new new problems

Learning Process LP

- Learning Process is a **non trivial** process for identification of :

  Valid (tested)

  New

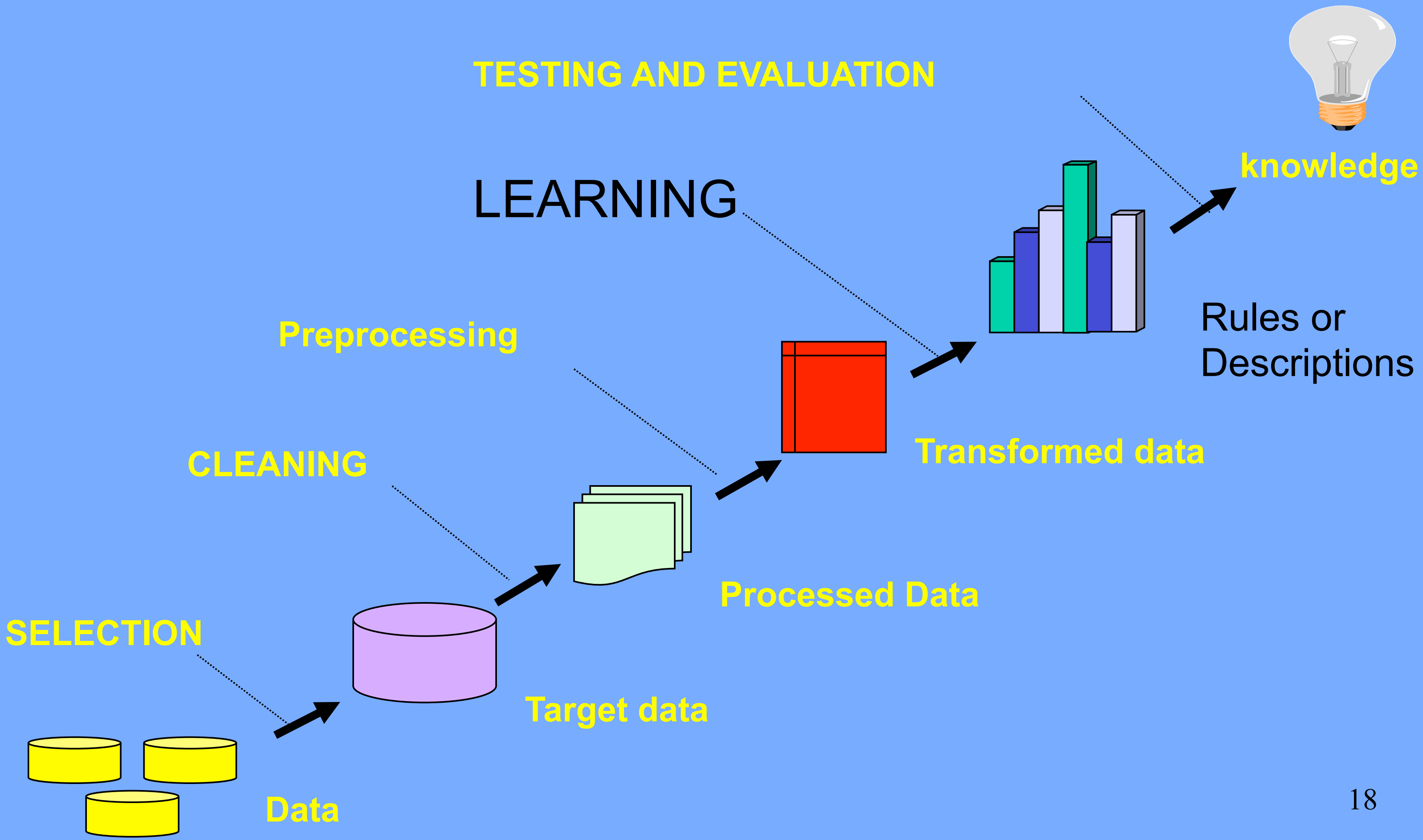
  Potentially useful

  Understable (when possible)

  patterns in data

  We talk about the Learning Process – but the same applies to Data Mining Process (as in our Book)

The Learning Process (LP)



TESTING AND EVALUATION

knowledge

LEARNING

Rules or Descriptions

Preprocessing

Transformed data

CLEANING

Processed Data

SELECTION

Target data

Data

LEARNING

- **Learning** is **a step** of the LP process in which algorithms are applied to look for **patterns in data**

- It is necessary to TEST and EVALUATE obtained **patterns**

- It is also necessary to apply first the **preprocessing** operation;
- to clean and preprocess the data in order to obtain **significant patterns**

Steps of the Learning Process

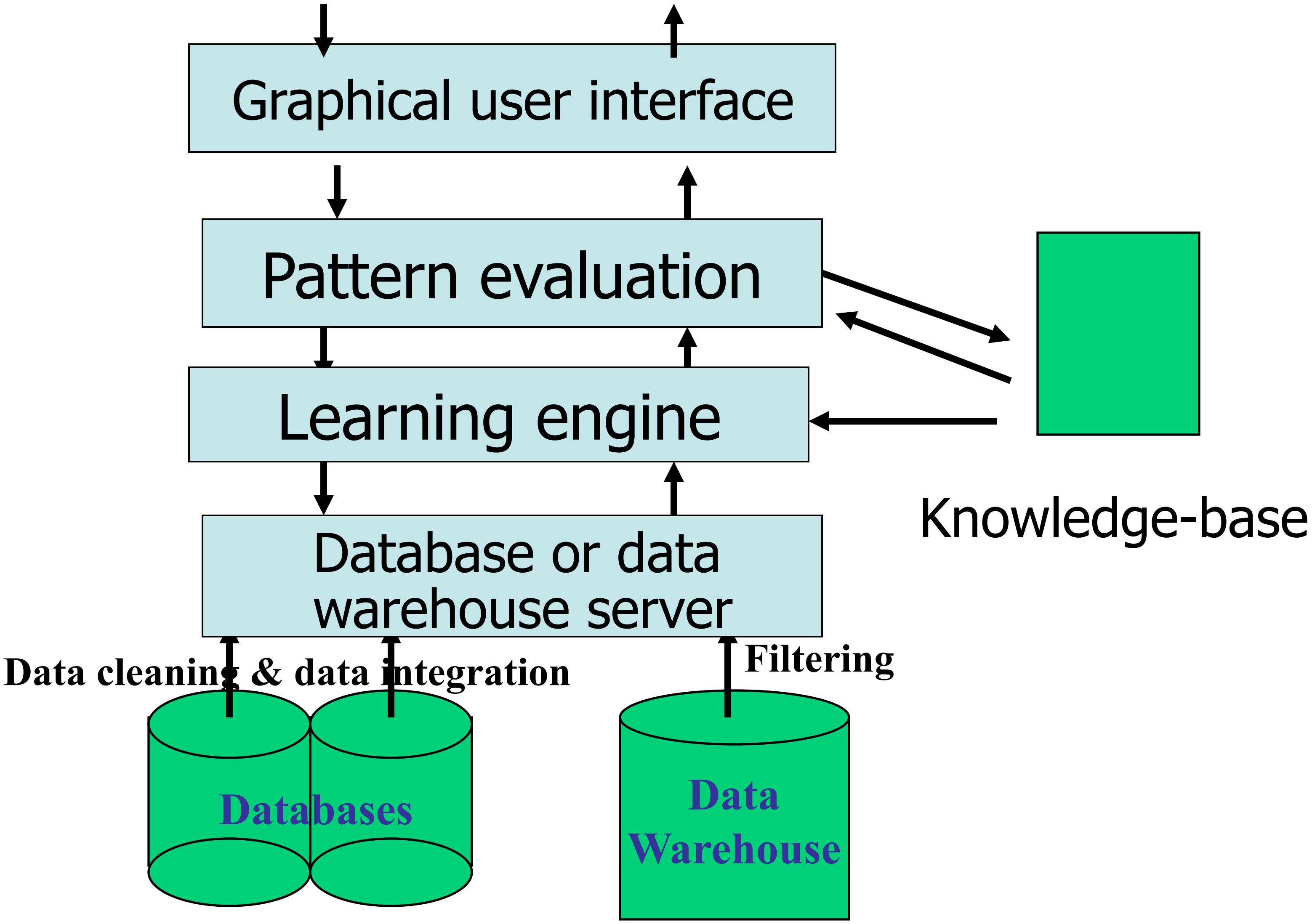
Preprocessing: includes all the operations that have to be performed **before** a learning algorithm is applied

Training: algorithms are applied to training data in order to obtain (**learn**) the patterns

Testing: testing methods are applied to **test** the learned patterns

Interpretation: discovered patterns are presented in a proper format and the user decides if it is neccesary **to re-iterate the algorthms**

Architecture of a Typical Learning System



Learning
What Kind of Data?

- Relational Databases
- Data warehouses (Data Mining)
- Transactional databases
- Advanced DB and information repositories

  Object-oriented and object-relational databases

  Spatial databases

  Time-series data and temporal data

  Text databases and multimedia databases

  Heterogeneous and legacy databases

  WWW

RELATIONAL DATA

- **We assume** for our considerations that **data used** in the learning algorithms are presented in a form of a **relational table** with the **key attribute removed**

Learning the **Characteristic Rules**

- ***It is a process which aim is to find rules that describe characteristic properties of a concept. They take the form***

***If** concept **then** characteristics*

- C=1 → A=1 & B=3 **25%** (support: there are 25% o the records for which the rule is true)
- C=1 → A=1 & B=4 **17%**
- C=1 → A=0 & B=2 **16%**

Learing the **Discriminat Rules**

- ***It is a process which aim is to find rules that allow us to discriminate the objects (records) belonging to a given concept (one class) from the rest of records (classes)***

If characteristics then concept

- **A=0 & B=1 → C=1   33% 83% (support, confidence: the conditional probability of the concept given the characteristics)**
- **A=2 & B=0 → C=1   27% 80%**
- **A=1 & B=1 → C=1   12% 76%**
- Discriminant rules can be accepted even if they have a low support (and high confidence)

Learning Functionalities

- **Classification, Classification Prediction** **is also called** **Supervised Learning**
- **Supervised Learning**

  Finding models (rules) that describe (characterize) or/ and distinguish (discriminate) classes or concepts for future prediction

  Example: classify countries based on climate (characteristics), or classify cars based on gas mileage and use it to predict classification of a new car

  Models, algorithms, methods: decision-tree, neural network, Bayes Network, Rough Sets, genetic algorithms

  Presentation of results: characteristic and /or discriminant rules- converged neural network, or Bayes network

Clustering
Unsupervised Learning

- **Cluster analysis (statistical method)**

  Class label is unknown;

  algorithms group data to form new classes;

  It is also called **unsupervised learning**

  For example: cluster houses to find distribution patterns

  Clustering is based on the principle:

  maximizing the intra-class similarity and **minimizing** the interclass similarity

Clustering

- Database segmentation

- Given a set of objects (records) the algorithm obtains a division of the objects into clusters in which the distance of objects inside a claster is minimal and the distance among objects of diferent clusters is maximal

- **Unsupervised learning**

Classification
Supervised Learning

- **Given** a set of objects a **classification algorithm** builds a set of **discriminant** and /or **characterization rules**

- or **other descriptions** in order to be able, as the next step, to **classify** unknown sets of objects

- This is also called a **supervised learning**

Classification Methods, Models, Algorithms

- DESCRIPTIVE:
  Decision Trees (ID3, C4.5)
- Rough Sets
- Genetic Algorithms
- STATISTICAL:
- Neural Networks
- Bayesian Networks

Summary

- Learning: discovering interesting patterns from often large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- Learning process LP includes data cleaning, data integration, data selection, transformation, training, testing, pattern evaluation, and knowledge presentation
- Learning can be performed in a variety of information repositories

Preprocesing

Preprocesing

- Preprocessing is a process in which we
- select, integrate, and clean the data;
- decide which kind of patterns are needed;
- decide which algorithm is the best;
- prepare data for algorithms

Implementation Preparation (1)

- Identify the problem to be solved.
- Study it in detail
- Explore the solution space,
- Find one acceptable solution (feasible to implement)
- Specify the solution
- Prepare and preprocess the data

Preparation

- Remember GIGO! (garbage in gabage out)
- Add some data, if necessary
- Structure the data in a proper form
- Be careful with incomplete and noisy data

Studying the data

- The surrounding world consists of objects (data) and the **Learning Process** goal is to find the **relationships among objects**

- The objects are characterized by properties (attributes, values of attributes) that have to be analized

- The results (rules, other descriptions) are **valid (true)** under certain circumstances (data we learn from) and in certain moments (avaible data at the moment)

Types of data

- Generally we distinguish:

  Quantitative Data

  Qualitative Data

- Bivaluated: often very useful
- Null Values are not applicable
- Missing data usually not acceptable

What to take into account

- Eliminate redundant records
- Eliminate out of range values of attributes
- **Decide a generalization level**
- Decide on the accuracy level

Summary

- The preprocessing is usually required and is an essential part of the **LP process**
- If preprocessing is not well performed the patterns obtained could be of no use
- Preprocessing is a tedious task that could even take more time that the **Learning proper**