Knowledge Discovery in Databases

Javier Béjar

CS - MIA

AMLT - 2016/2017
1. Knowledge Discovery in Databases
   - Introduction
   - Definitions of KDD

2. The KDD process
   - Steps of KDD
   - Discovery goals
   - Mining Methodologies

3. Applications

4. Tools

5. Challenges
1 Knowledge Discovery in Databases

2 The KDD process

3 Applications

4 Tools

5 Challenges
Knowledge Discovery in Databases

- Practical application of the methodologies from machine learning/statistics to large amounts of data
- The main problem addressed is the impossible task of manually analyzing (make sense of) all the data we are systematically collecting
- These methodologies are useful for automating/helping the process of analysis/discovery
- The final goal is to extract (semi)automatically actionable/useful knowledge

“We are drowning in information and starving for knowledge”
The high point of KDD starts around late 1990s

Many companies show their interest in obtaining the (possibly) valuable information stored in their databases (purchase transactions, e-commerce, web data, ...) 

The goal is to obtain information that can lead to better commercial strategies and practices from a better understanding of the consumers preferences and their behaviour

Many companies are putting a lot of effort on the development/use of this kind of technology (analysis and tools)

Several buzz words have appeared: Business Intelligence, Business Analytics, Predictive Analytics, Data Science, Big Data ...
Not only business data are in need of these kinds of techniques.

Analyzing scientific data has supposed an important impulse:
- Space probes
- Remote sensors on satellites
- Astronomical observations (big array observatories)
- Large scientific experiments (LHC, ITER)
- Genome Project, microarray data \(\Rightarrow\) Bioinformatics
- Neuroscience (Human Brain Project)

Data grows faster than the ability to analyze it.
KDD: Machine learning

- **Inductive machine learning**: Discovery of patterns/models from data
- Supervised discovery/Unsupervised discovery
- Unstructured/Structured representations
- Logic representations/Probabilistical representations
- Scalability
KDD: Statistics/Data Analysis

- **Statistical Data Modeling**: Fitting of probability models to data
- Supervised/Unsupervised modeling
- Structured models
- Probabilistic representation/interpretation of data
- Scalability
- Statistical Machine Learning
KDD: Databases/Algorithmics/Visualization

- Data access:
  - SQL vs NoSQL
  - Distributed file systems
  - Redundancy/Fault Tolerance/Parallelism

- Databases for structured data: Transactions, Graphs, Time sequences

- Distributed processing paradigms/scalability: MapReduce, Hadoop, Spark, ..

- Data visualization: from data cubes to structured data representation
KDD definitions

“It is the search of valuable information in great volumes of data”

“It is the explorations and analysis, by automatic or semiautomatic tools, of great volumes of data in order to discover patterns and rules”

“It is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”
Elements of KDD

Pattern: Any representation formalism capable to describe the common characteristics of a group if instances

Valid: A pattern is valid if it is able to predict the behaviour of new information with a degree of certainty

Novelty: It is novel any knowledge that it is not know respect the domain knowledge and any previous discovered knowledge

Useful: New knowledge is useful if it allows to perform actions that yield some benefit given a established criteria

Understandable: The knowledge discovered must be analyzed by an expert in the domain, in consequence the interpretability of the result is important
1. Knowledge Discovery in Databases

2. The KDD process

3. Applications

4. Tools

5. Challenges
The KDD process

KDD as a process

- The actual discovery of patterns is only one part of a more complex process
- Raw data is not always ready for processing (80/20 project effort)
- Some general methodologies have been defined for the whole process (CRISP-DM or SEMMA)
- These methodologies address KDD as an engineering process, despite being business oriented are general enough to be applied on any data discovery domain
The KDD process (I)

Steps of the Knowledge Discovery in DB process

1. Domain study
2. Creating the dataset
3. Data preprocessing
4. Dimensionality reduction
5. Selection of the discovery goal
6. Selection of the adequate methodologies
7. Data Mining
8. Result assessment and interpretation
9. Using the knowledge
The KDD process (II)

1. Study of the domain
   Gather information about the domain. Characteristics, goal of the discovering process (attributes, representative examples, types of patterns, sources of data)

2. Creating the dataset
   From the information of the previous step it is decided what source of data will be used. It has to be decided what attributes will describe the data and what examples are needed for the goals of the discovery process
3. **Data preprocessing and cleaning**
   It has to be studied the circumstances that affect the quality of the data
   - Outliers
   - Noise (does it exists?, does it present any pattern?, can it be reduced?)
   - Missing values
   - Discretization of continuous values
4. **Data reduction and projection**

We have to study what attributes are relevant to our goal (depending on the task some techniques can be used to measure the relevance of the attributes) and the number of examples that are needed. Not all the data mining algorithms are scalable

- Instance selection (do we need all the examples? sampling techniques)
- Attribute selection (what is really relevant?)
The KDD process (VI) - Attribute selection

- It is very important to use methods for attribute selection:
  - Reduces the dimensionality of the data (curse of dimensionality)
  - Eliminates/Reduces irrelevant and redundant information
  - The result of the process is easier to interpret

- Attribute selection techniques:
  - Mathematical/Statistical techniques: Principal component analysis (PCA), projection pursuit, Multidimensional scaling
  - Heuristics for attribute relevance evaluation (ranking of attributes, search in the space of subsets of attributes)
5. **Selecting the discovery goal**
   The characteristics of the data, the domain and the aim of the project determines what kind of analysis are feasible or possible (group partitioning, summarization, classification, discovery of attribute relations, ...)

6. **Selecting the adequate methodologies**
   The goal and the characteristics of the data determines the more adequate methodologies
The KDD process (VIII)

7. Applying the methodologies (Data Mining)
The different parameters of the chosen methodologies has to be adjusted by experimentation and analysis in order to obtain the best possible results.

8. Interpreting the results
From the knowledge of the domain (expert) it will be assessed the relevance and importance of the result. This interpretation step could suppose feedback for the previous steps, it is possible that some adjustments are needed or some previous decisions have to be changed.

9. Incorporating the new knowledge
The new knowledge is used to perform the intended task goal of the discovery process.
Goals of the KDD process

There are different goals that can be pursued as the result of the discovery process, among them:

**Classification:** We need models that allow to discriminate instances that belong to a previously known set of groups (the model could or could not be interpretable)

**Clustering/Partitioning/Segmentation:** We need to discover models that clusters the data into groups with common characteristics (a characterizations of the groups is desirable)

**Regression:** We look for models that predicts the behaviour of continuous variables as a function of others
Goals of the KDD process

Summarization: We look for a compact description that summarizes the characteristics of the data

Causal dependence: We need models that reveal the causal dependence among the variables and assess the strength of this dependence

Structure dependence: We need models that reveal patterns among the relations that describe the structure of the data

Change: We need models that discover patterns in data that has temporal or spatial dependence
Methodologies for KDD

There are a lot of methodologies that can be applied in the discovery process, the more usual are:

- **Decision trees, decision rules:**
  - Usually are interpretable models
  - Can be used for: Classification, regression, and summarization
  - trees: C4.5, CART, QUEST, rules: RIPPER, CN2, ..

- **Classifiers, Regression:**
  - Low interpretability but good accuracy
  - Can be used for: Classification and regression
  - Statistical regression, function approximation, Neural networks, Support Vector Machines, k-NN, Local Weighted Regression, ..
Methodologies for KDD

- **Clustering:**
  - Its goal is to partition datasets or discover groups
  - Can be used for: Clustering, summarization
  - Statistical Clustering, Unsupervised Machine learning, Unsupervised Neural networks (Self-Organizing Maps)

- **Dependency models (attribute dependence, temporal dependence, graph substructures):**
  - Its goal is to obtain models (some interpretables) of the dependence relations (structural, causal temporal) among attributes/instances
  - Can be used for: causal dependence discovery, temporal change, substructure discovery
  - Bayesian networks, association rules, Markov models, graph algorithms, ...
1. Knowledge Discovery in Databases

2. The KDD process

3. Applications

4. Tools

5. Challenges
Applications

- **Business:**
  - Costumer segmentation, costumer profiling, costumer transaction data, customer churn
  - Fraud detection
  - Control/analysis of industrial processes
  - e-commerce, on-line recommendation
  - Financial data (stock market analysis)

- **WEB mining**
  - Text mining, document search/organization
  - Social networks analysis
  - User behavior
Scientific applications:
- Medicine (patient data, MRI scans, ECG, EEG, ...)
- Pharmacology (Drug discovery, screening, in-silicon testing)
- Astronomy (astronomical bodies identification)
- Genetics (gen identification, DNA microarrays, bioinformatics)
- Satellite/Probe data (meteorology, astronomy, geological, ...)
- Large scientific experiments (CERN LHC, ITER)
1. Knowledge Discovery in Databases

2. The KDD process

3. Applications

4. Tools

5. Challenges
There are a lot of tools available for KDD

Some tools were developed at universities (C5.0, CART/MARS) and have become a commercial product, others still remain open source (Weka, R, scikit-learn)

Big fish eats little fish (C5.0 → Clementine → SPSS-clementine → IBM DBMiner)

Data analysis software companies incorporate KDD techniques inside classical data analysis tools (SPSS, SAS)

Companies selling databases add KDD tools as an added value (IBM DB2 (intelligent Miner), SQL Server, Oracle)

Machine Learning as a Service (Amazon, Microsoft, Google, IBM Watson, Big ML, ...)
Tools for the course

- **Python**
  - General programming language, easy to learn
  - numpy, scipy, pandas
  - scikit-learn (http://scikit-learn.org)
  - Data preprocessing, Clustering Algorithms, Association Rules, ...

- **R** (http://cran.r-project.org/)
  - Statistic analysis oriented language, more steep learning curve
  - Many packages
  - Data preprocessing, Clustering Algorithms, Association Rules, ...
1. Knowledge Discovery in Databases
2. The KDD process
3. Applications
4. Tools
5. Challenges
Open problems

- Scalability (More data, more attributes)
- Overfitting (Patterns with low interest)
- Statistical significance of the results
- Methods for temporal data/relational data/structured data
- Methods for data cleaning (Missing data and noise)
- Pattern comprehensibility
- Use of domain knowledge
- Integration with other techniques (OLAP, Data Warehousing, Business Intelligence, Intelligent Decision Support Systems)
- Privacy