Machine Learning and Data Mining

Academic Year 2023 – 2024
Alma Mater Studiorum · University of Bologna

Contents

Acronyms

BI Business Intelligence

DM Data Mart

DSS Decision Support System

DWH Data Warehouse

EIS Executive Information System

ERP Enterprise Resource Planning

ETL Extraction, Transformation, Loading

MIS Management Information System

OLAP Online Analysical Processing

OLTP Online Transaction Processing

1 Introduction

1.1 Data

Data Collection of raw values. Data

Information Organized data (e.g. relationships, context, ...). Information

Knowledge Understanding information. Knowledge

1.1.1 Data sources

Transaction Business event that generates or modifies data in an information system (e.g. database). Transaction

Signal Measure produced by a sensor. Signal

External subjects

1.1.2 Software

Online Transaction Processing (OLTP) Class of programs to support transaction oriented applications and data storage. Suitable for real-time applications. Online Transaction Processing

Enterprise Resource Planning (ERP) Integrated system to manage all the processes of a business. Uses a shared database for all applications. Suitable for real-time applications. Enterprise Resource Planning

1.1.3 Insight

Decision can be classified as:

Structured Established and well understood situations. What is needed is known. Structured decision

Unstructured Unplanned and unclear situations. What is needed for the decision is unknown. Unstructured decision

Different levels of insight can be extracted by:

Management Information System (MIS) Standardized reporting system built on existing OLTP. Used for structured decisions. Management Information System

Decision Support System (DSS) Analytical system to provide support for unstructured decisions. Decision Support System

Executive Information System (EIS) Formulate high level decisions that impact the organization. Executive Information System

Online Analysical Processing (OLAP) Grouped analysis of multidimensional data. Involves large amount of data. Online Analysical Processing

Business Intelligence (BI) Applications, infrastructure, tools and best practices to analyze information. Business Intelligence

Big data Large and/or complex and/or fast changing collection of data that traditional DBMSs are unable to process. Big data

Structured e.g. relational tables.

Unstructured e.g. videos.

Semi-structured e.g. JSON.

Anaylitics Structured decision driven by data. Anaylitics

Data mining Discovery process for unstructured decisions. Data mining

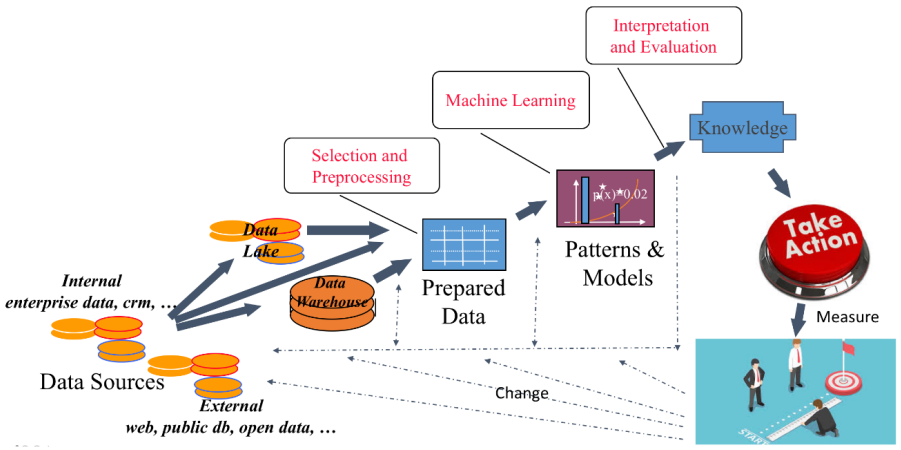


Figure 1.1: Data mining process

Machine learning Learning models and algorithms that allow to extract patterns from data. Machine learning

2 Business Intelligence

Business Intelligence Transform raw data into information. Deliver the right information to the right people at the right time through the right channel. — Business Intelligence

Data Warehouse (DWH) Optimized repository that stores information for decision making processes. DWHs are a specific type of DSS. — Data Warehouse

Features:

- Subject-oriented: focused on enterprise specific concepts.
- Integrates data from different sources and provides an unified view.
- Non-volatile storage with change tracking.

Data Mart (DM) Subset of the primary DWH with information relevant to a specific business area. — Data Mart

2.1 Online Analysical Processing (OLAP)

OLAP analyses Interactively navigate the information in a data warehouse. Allows to visualize different levels of aggregation. — Online Analysical Processing (OLAP)

OLAP session Navigation path created by the operations of a user.

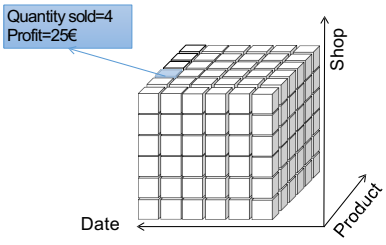


Figure 2.1: OLAP data cube

2.1.1 Operators

Roll-up Increases the level of aggregation (i.e. `GROUP BY` in SQL). Some details are collapsed together. — Roll-up

Drill-down Reduces the level of aggregation. Some details are reintroduced. — Drill-down

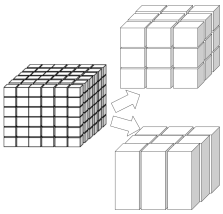
Slide-and-dice The slice operator reduces the number of dimensions (i.e. drops columns). — Slide-and-dice

The dice operator reduces the number of data being analyzed (i.e. `LIMIT` in SQL).

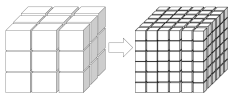
Pivot Changes the layout of the data to analyze it from a different viewpoint. — Pivot

Drill-across Links concepts from different data sources (i.e. `JOIN` in SQL). Drill-across

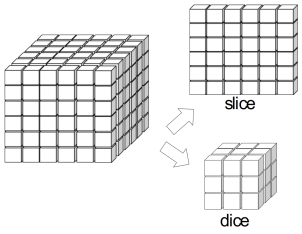
Drill-through Switches from multidimensional aggregated data to operational data (e.g. a spreadsheet). Drill-through



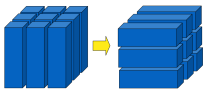
(a) OLAP roll-up



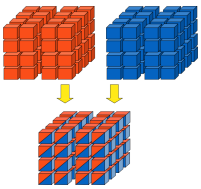
(b) OLAP drill-down



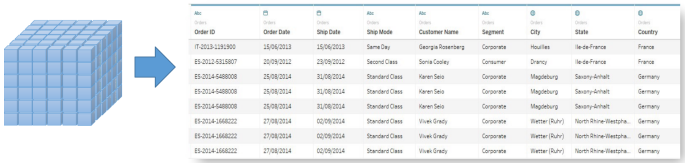
(c) OLAP slide-and-dice



(d) OLAP pivot



(e) OLAP drill-across



(f) OLAP drill-through

2.2 Extraction, Transformation, Loading (ETL)

The ETL process extracts, integrates and cleans operational data that will be loaded into a data warehouse. Extraction, Transformation, Loading (ETL)

2.2.1 Extraction

Extracted operational data can be:

Structured with a predefined data model (e.g. relational DB, CSV) Strucured data

Untructured without a predefined data model (e.g. social media content) Unstrucured data

Extraction can be of two types:

Static The entirety of the operational data are extracted to populate the data warehouse for the first time. Static extraction

Incremental Only changes applied since the last extraction are considered. Can be based on a timestamp or a trigger. Incremental extraction

2.2.2 Cleaning

Operational data may contain:

Duplicate data

Missing data

Improper use of fields (e.g. saving the phone number in the `notes` field)

Wrong values (e.g. 30th of February)

Inconsistency (e.g. use of different abbreviations)

Typos

Methods to increase the quality of the data are:

Dictionary-based techniques Lookup tables to substitute abbreviations, synonyms or typos. Applicable if the domain is known and limited. — Dictionary-based cleaning

Approximate merging Merging data that do not have a common key. — Approximate merging

- **Approximate join** Use non-key attributes to join two tables (e.g. using the name and surname instead of an identifier).
- **Similarity approach** Use similarity functions (e.g. edit distance) to merge multiple instances of the same information (e.g. typo in customer surname).

Ad-hoc algorithms — Ad-hoc algorithms

2.2.3 Transformation

Data are transformed to respect the format of the data warehouse:

Conversion modifications of types and formats (e.g. date format) — Conversion

Enrichment creating new information by using existing attributes (e.g. compute profit from receipts and expenses) — Enrichment

Separation and concatenation Denormalization of the data: introduces redundances (i.e. breaks normal form[1]) to speed up operations. — Separation and concatenation

2.2.4 Loading

Adding data into a data warehouse:

Refresh The entire DWH is rewritten. — Refresh loading

Update Only the changes are added to the DWH. Old data is not modified. — Update loading

[1] `https://en.wikipedia.org/wiki/Database_normalization`