Pattern mining and GALACTIC

The Galactic Organization <contact@thegalactic.org>

2018-2022  [1]

Pattern Mining

- Pattern mining is a clustering approach aiming at generating a hierarchy of patterns.
- Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold.
- Pattern mining helps in mining association and correlations among data.
- Pattern mining has been introduced by the APriori algorithm for symbolic data (1994), then restricted to closed itemsets (1999), and extended to structured data (sequences, graphs, ...)
- But a deluge of patterns is produced

Main Principle

Level by level generation

- Initialization with *<all the objects - common pattern>*
- Iteration level by level:
  - Identification of *candidates* for the next level by addition of an atomic information to the common patterns
  - Checking the validity of each *candidate*
    - strictly decreasing support, without doublings, without gap of level
  - For each valid candidate:
    - <objects sharing the candidate - common patterns >

- Atomic information can be a new attributes (for symbolic data) - a new item (for sequences) - a node or an edge (for graphs) ...

The first main algorithms

	Symbolic data	Sequences	Graphs
Patterns	Itemsets (APriori, 1994)	Subsequences, Prefix (GSP, 1996 - Spade, 2001)	Subgraphs
Closed patterns	Closed itemsets (Pasquier, 1999)	Maximal subsequences (CloSpan, 2003)	Maximal subgraphs (GSpan, 2002)

Formal Concept Analysis (FCA)

- Formal Concept Analysis has been introduced by R. Wille (1982) to extract information from binary data under the form of a concept lattice
  - each concept is composed of data and their common attributes
- FCA is an application of lattice theory and the work of M. Barbut and B.M Monjardet (1972) with:
  - the structure of *concept lattice* or *closure lattice*
  - the *Galois connection* and the *closure operator* and
  - the possibility of extracting *basis of rules* with *minimal generators* as premises
- FCA has been extended to non binary data with pattern structures (2001) when a Galois connection exists between objects and their description by common patterns
- Abstract Conceptual Navigation (ACN) is a user driven method by navigation inside the lattice.

Bordat's algorithm

Bordat's algorithm

- Input: a binary table (context), and its closure operator φ
- Compute the minimal closure $\varphi(\emptyset)$
  - The minimal concept <all the objects - $\varphi(\emptyset)$>
- Recursive computation of the immediate successor of each closure:
  - The immediate successors of a closure F are the inclusion minimal set of the family $\{\varphi(x + F)\}$ (Bordat's theorem)

The NextPriorityConcept algorithm

NextPriorityConcept is a new pattern mining algorithm (2021) for mining classical and structured data

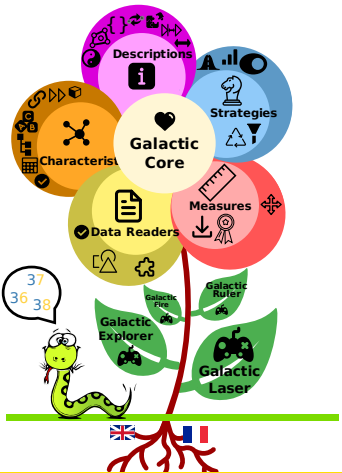
- NPC is issued from FCA and inspired from Bordat's algorithm:
  - A concept is a pair <subgroups - patterns>
  - The lattice is generated level by level
  - The lattice structure is maintained with a mechanism of constrainsts propagation
- Patterns are monadic predicates issued from *generic description* of data:
  - For any type od data (classical, structured) that can be mixed
- Candidates at each level are generating by predicates issued from *generic strategies*
  - Classical approaches with strategies generating all the possible candidates
  - Data discovery with more sophisticated strategies
  - Strategies can be mixed with filter measures for an interactive user driven approach

GALACTIC

Written in python, Fully extensible

The GALACTIC framework is architecturally designed with:

- a core library with the NextPriorityConcept algorithm
- characteristic plugins
- description plugins
- strategy plugins
- measure plugins
- data reader plugins
- localization plugins
- applications



Current scientific issues in AI

1. Need for explicability and legibility (legal and societal context)
   - *Black box* vs *White box*
2. Handle complex data (sequences, graphs, temporal information, ...)
   - *Embedding* vs *Keeping the structure*
3. Difficulty in generating a ground truth
   - *Learning* vs *Clustering*
4. "Responsible digital" approaches
   - *Convergence* vs *One-pass*
5. Process huge data
   - *Data Mining* vs *Data Discovery*

Black box

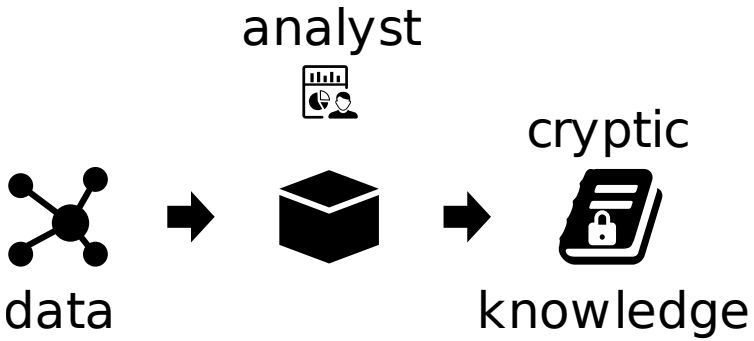


Figure 1: Deep learning

White box

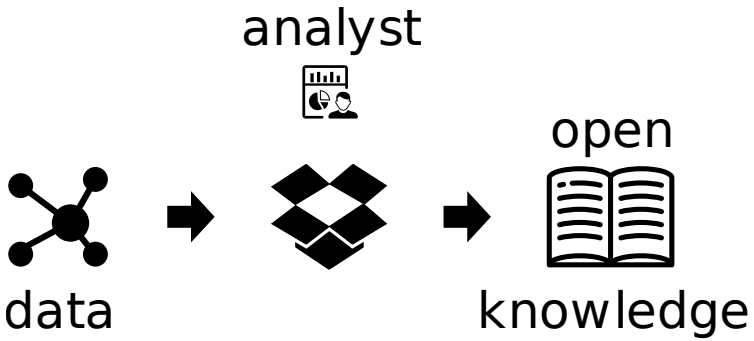


Figure 2: Classical Formal Concept Analysis

Interactive white box

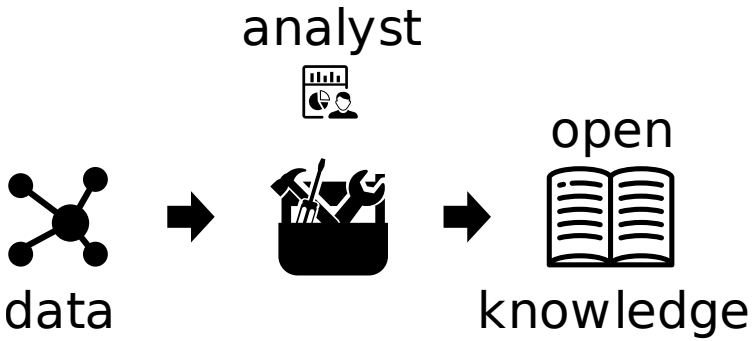


Figure 3: User driven Formal Concept Analysis

Conclusion

	Explicability	Complex data	Huge data
Pattern Mining	White box	Specific methods	Data mining
FCA	White box	Specific methods	Data mining
GALACTIC	Interactive White box	Generic method	Data mining & discovery