LINFO2364 Mining Patterns in Data

Issambre L'Hermite Dumont

This summary may not be up-to-date, the newer version is available at this address:
https://github.com/SimonDesmidt/Syntheses

Academic year 2025-2026 - Q2

UCLouvain

Contents

Introduction

In our data-driven world, the ability to extract meaningful information from vast datasets is crucial. Understanding all the aspect of this discipline is essential to derive those meaningful information, and this is the goal of this course. First, we need to define some key concepts.

1.1 Definitions

Definition 1.1. A pattern is a recurring structure in a dataset.

Patterns can be simple or complex, relevant or irrelevant. Their advantages is that they are interpretable. When found, relevant patterns can be used to make predictions, to understand the underlying structure of the data, and to make informed decisions.

Definition 1.2. Data mining is the process of discovering interesting patterns, models, and other kinds of knowledge in large data sets.

1.1.1 Type of data

We can mine data out of various types of structure of data:

- **Tabular data**: Data is organized in rows and columns. Example: spreadsheets, databases.
- **Sequences**: Data points are ordered in a sequence. Example: DNA sequences, text data.
- **Graphs, trees, networks**: Data is represented as nodes and edges. Example: social networks, web graphs.

Those structures can be discretes, continuous, enumerable data, etc. Those structures, can be combined to form more complex data types. And they can be highly structured, semi-structured, or unstructured.

Definition 1.3. Highly structured data are relational databases, with uniform record or table-like structures, with a fixed set of well-defined attributes. This is rarely the case in real-world data.

Definition 1.4. Semi-structured data are not as structured as in relational databases, but presents some structure with clearly defined semantic meaning. For example:

- Transactional dataset: structured into transactions, but each transaction is an unstructured set of values
- Sequence data set: unstructured collection of ordered sequences of values
- Graphs: set of nodes connected by a set of edges, with edges labelled given some semantic

Definition 1.5. Unstructured data have no predefined structure or organization. For example: text documents, images, audio files, videos.

Those requires advanced techniques to extract patterns, like deep learning or domain-specific methods.