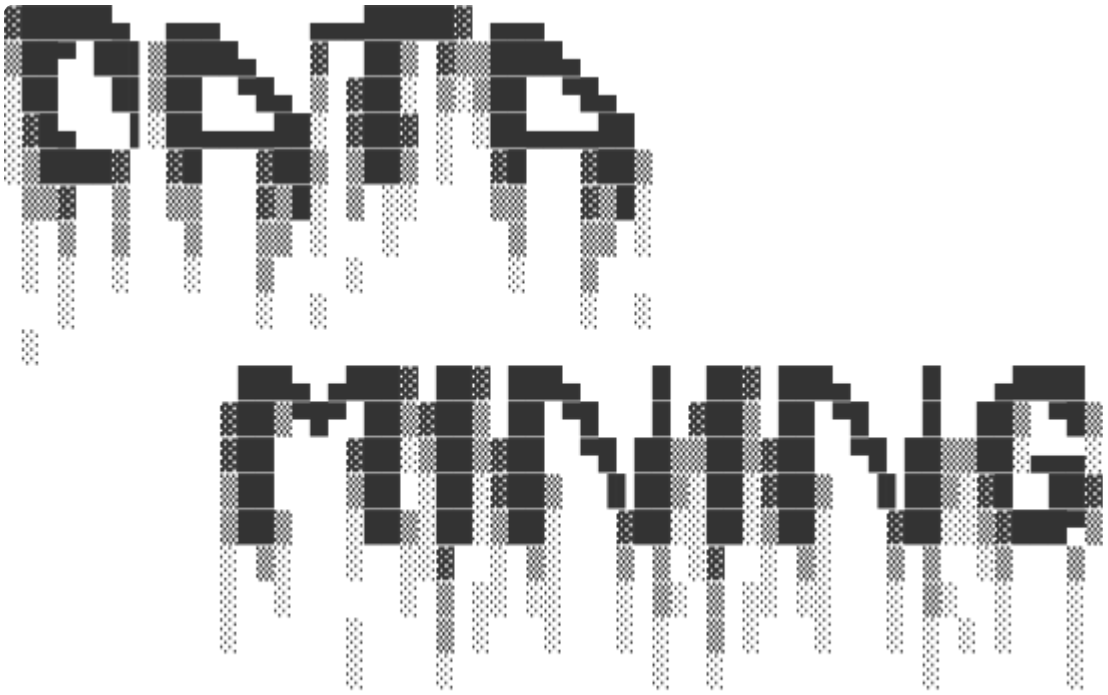


#HCMUS #DataMining

Thông tin kỳ thi

Thời gian: 9h55 - 29/06/2024

Phòng thi: E403 (NVC)

Overview

Đây là note của toàn bộ môn học. Chuẩn bị cho kỳ thi cuối kỳ sắp tới.

Note sẽ bám theo sách **Data Mining The Textbook** gồm các nội dung sau:

1. An Introduce to Data Mining
2. Data Preparation
3. Similarity & Distances
4. Association Pattern Mining
5. Cluster Analysis

6. Outlier Analysis
7. Data Classification

1. Introduce to Data Mining

1.1. Introduce

Data mining là quá trình collect, clean, process, analyzing và gain useful insights từ data của nhiều domain khác nhau.

Do đó, 'data mining' là cách gọi tổng quát để miêu tả các khía cạnh khác của data processing.

Các nhà phân tích dữ liệu sẽ sử dụng hệ thống xử lý, trong đó, các dữ liệu thô sẽ được:

- Thu thập
- Làm sạch
- Chuyển đổi thành một định dạng tiêu chuẩn hóa

Dữ liệu có thể có nhiều định dạng hoặc kiểu khác nhau.

- Định lượng (quantitative)
- Định tính (categorical || quanlitative)
- Văn bản (text)
- Không gian (spatial), thời gian (temporal)
- Biểu đồ (graph-oriented)

Hầu hết dữ liệu trong thực tế là multidimensional data. Các loại dữ liệu có cấu trúc phức tạp dần dần có tỷ lệ ngày càng tăng

1.2. The Data Mining Process

Data Mining Process là một pipeline chứa nhiều phases khác nhau như:

1. **Data Collection**: là quá trình yêu cầu sử dụng nhiều phương pháp khác nhau để collect dữ liệu. Đây là một quá trình rất quan trọng ảnh hưởng trực tiếp đến DMP (stand for Data Mining Process). Sau quá trình collection, data sẽ được lưu trữ ở database hoặc data warehouse (dành cho processing)
2. **Feature extraction và data cleaning**: Sau khi dữ liệu đã được collect, chúng thường không phù hợp để processing ngay. Trong nhiều trường hợp, data có nhiều type khác nhau sẽ mix lại với nhau, transform chúng, biến data thành dạng suitable type, chẳng hạn như multidimensional, time series, semistructured format. Thường quá trình feature extraction sẽ song song với data cleaning. Đầu ra cuối cùng sẽ là structured data set, có thể được sử dụng bởi computer program. Sau đấy, dữ liệu sẽ lại một lần nữa được lưu trữ ở database.
3. **Analytical processing and algorithms**: Đây là final part của DMP. Part này giúp đưa ra các useful insights hoặc các dự đoán có lợi

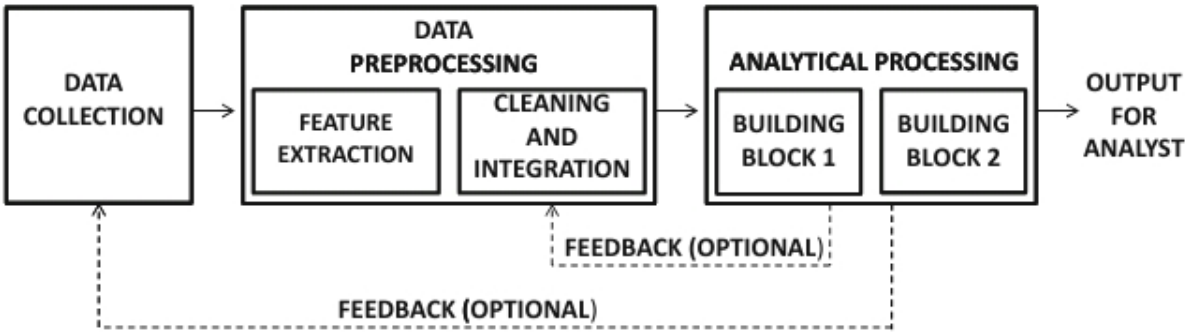


Figure 1.1: The data processing pipeline

1.2.1. The Data Preprocessing Phase

Đây là giai đoạn quan trọng nhất trong quá trình xử lý dữ liệu

Bao gồm các bước:

1. Feature Extraction
2. Data Cleaning
3. Feature selection and transformation

1.2.2. The Analytical Phase

1.3. The Basic Data Types

Có hai loại chính:

1.3.1. Dữ liệu định hướng không phụ thuộc

Là dạng dữ liệu đơn giản nhất và thường được gọi là dữ liệu nhiều chiều

Table 1.1: An example of a multidimensional data set

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Có hai loại phụ thuộc:

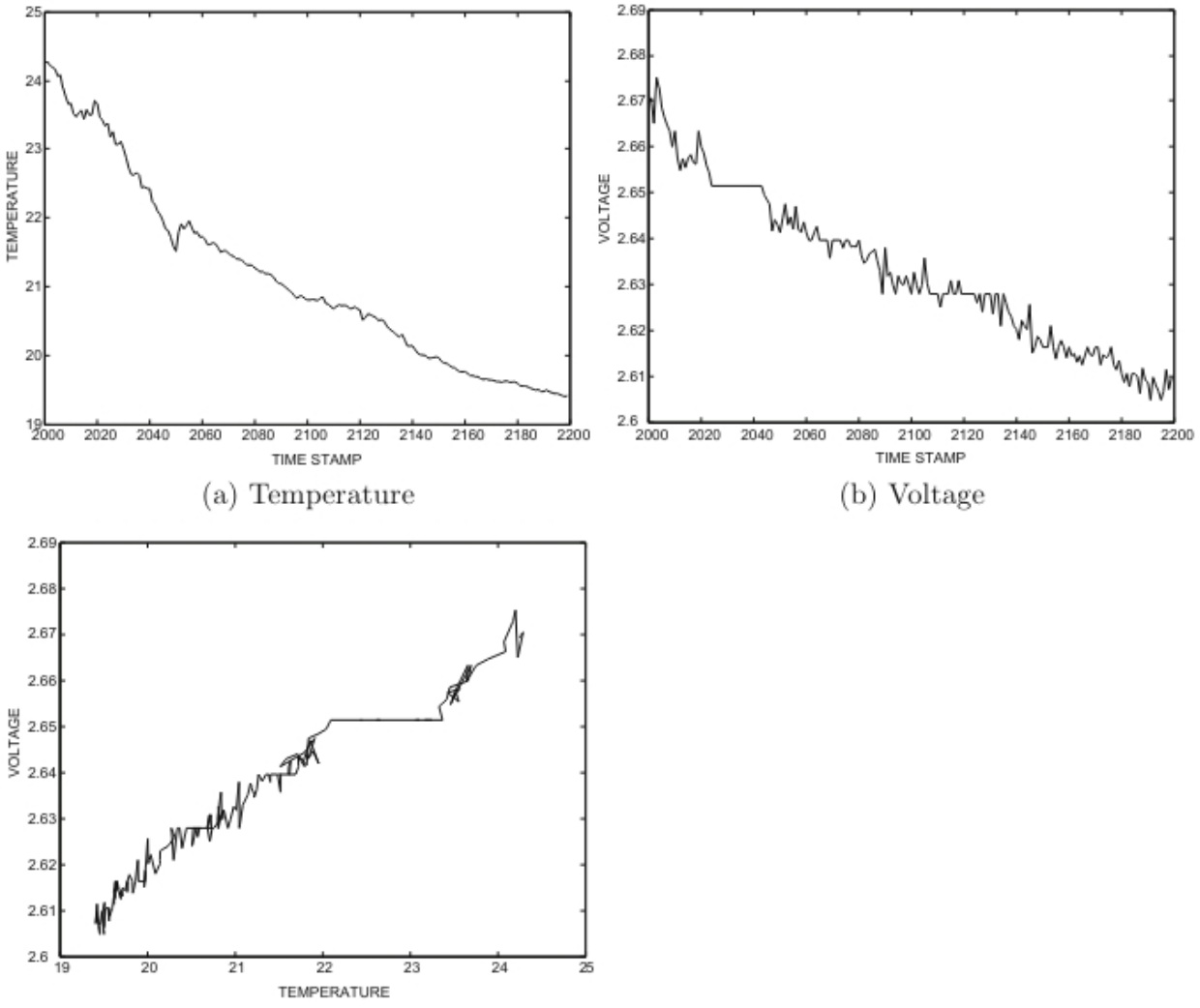
1. Phụ thuộc ngầm: các phụ thuộc giữa các mục dữ liệu không được chỉ định rõ ràng nhưng chúng "thường" tồn tại trong lĩnh vực đó. Ví dụ: nhiệt độ đo được từ cảm biến ở các thời điểm khác nhau, nếu ở hai thời điểm gần nhau mà nhiệt độ chêch lệch lớn thì đấy là dấu hiệu của sự không bình thường.
2. Phụ thuộc tường minh: Thường ám chỉ đến dữ liệu đồ thị (graph) hoặc mạng (network data) trong đó các cạnh được sử dụng để chỉ định mối quan hệ rõ ràng.

Dữ liệu định hướng phụ thuộc có thể được phân thành các loại sau:

- Time-Series data: Dữ liệu chuỗi thời gian chứa các giá trị thường được tạo ra bởi việc đo liên tục trong thời gian
- Discrete Sequences and Strings: Dãy rời rạc có thể được coi là biến thể của dữ liệu chuỗi thời gian. Giống như dữ liệu chuỗi thời gian, contextual attribute (thuộc tính ngữ cảnh) là một time stamp hoặc position index. Behavioral attribute (thuộc tính hành

vi) là một catogorical value. Do đó, dữ liệu dãy rời rạc được định nghĩa tương tự dữ liệu chuỗi thời gian.

- Spatial Data: bao gồm nhiều attribute không tuân theo không gian (nonspatial attributes) (ví dụ: nhiệt độ, áp suất...) được đo dựa trên không gian của chúng. Ví dụ: nhiệt độ bề mặt biển.
- Network and Graph Data:



(a) Temperature

(b) Voltage

(c) Temperature-voltage trajectory

Figure 1.2: Mapping of multivariate time series to trajectory data

1.3.2. Dữ liệu định hướng phụ thuộc

Được phân thành các loại như:

- Quantitative Multidimensional Data (dữ liệu nhiều chiều định lượng được)

- Categorical and Mixed Attribute Data (dữ liệu phân loại và dữ liệu thuộc tính trộn lẫn)
- Binary and Set Data (dữ liệu nhị phân và dữ liệu tập hợp)
- Text Data (dữ liệu văn bản)

Thông tin về sự phụ thuộc có từ trước có ảnh hưởng rất lớn đến DMP

1.4. The Major Building Blocks: A Bird’s Eye View