Data Mining Theory 5: Data Preparation and Preprocessing

Sebastião Jerónimo

October 2025

1 Introduction

Data preparation is a crucial step in the data mining process that involves transforming raw data into a format suitable for analysis and modeling. This process ensures that the quality of models improves (or at least remains the same) after preparation. Good data is a prerequisite for good models.

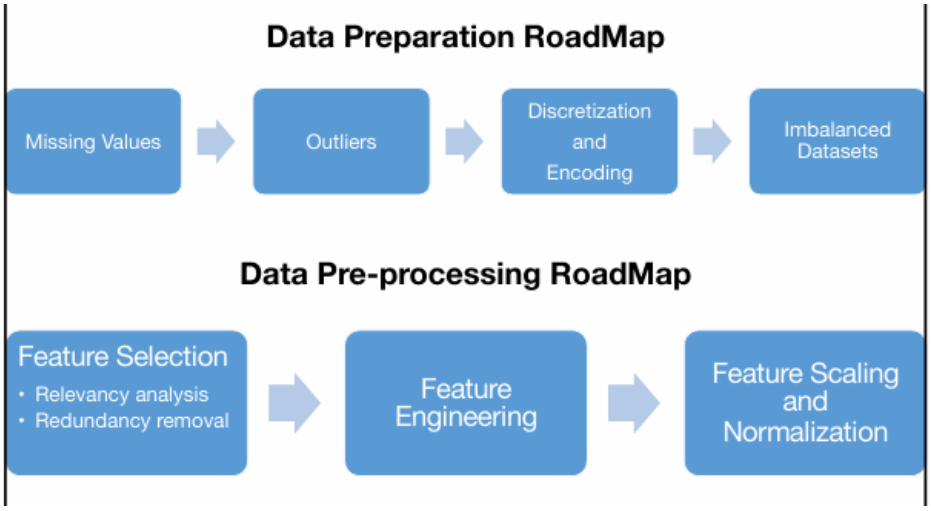

Figure 1: Data Preparation Roadmap

2 Objectives of Data Preparation

- Transform datasets to best expose their information content
- Improve model quality (or maintain it) after preparation
- Provide good data as a prerequisite for good models
- Apply theoretically-based techniques and experience-driven methods

3 Common Data Issues

Real-world data often suffers from several problems:

3.1 Incomplete Data

- Missing values
- Lacking attributes of interest
- Inappropriate levels of aggregation

3.2 Noisy Data

- Errors in measurement
- Outliers and anomalies
- External disturbances in information flow

3.3 Inconsistent Data

- Contradictory values (e.g., Age=42, Birthday=31/07/1997)
- Changes in scales over time
- Duplicate records with different values

4 Missing Data Treatment

4.1 Methods for Handling Missing Values

1. **Delete variables** - but this results in information loss
2. **Delete records** - may introduce potential bias
3. **Manual entry** - tedious and often infeasible
4. **Central tendency imputation** - using mean, median, or mode
5. **Subset-based imputation** - using central tendency within subgroups (e.g., men vs. women)
6. **Nearest neighbors imputation** - using values from similar individuals
7. **Predictive modeling** - using linear regression or multiple linear regression
8. **Explicit coding** - creating special codes for missing data

4.2 Practical Approach

- Start with the quickest and simplest option
- Compare model performance between full samples and those with estimated values
- If error is significantly higher, try alternative methods

5 Outlier Treatment

5.1 Definition and Causes

In statistics, an outlier is an observation point distant from other observations. Outliers may come from:

- Unusual but correct situations (the "Bill Gates effect")
- Incorrect measurements
- Errors in data collection
- Lack of coding for missing data

5.2 Detection Methods

- **Automatic thresholding** - imposing maximum and minimum values
- **Statistical methods** - using 3 ± average rule
- **Visual inspection** - examining histograms for:
  - Isolated bars separated from main distribution
  - Long tails stretching farther than expected
  - Bins with very few observations far from center
  - Asymmetric shapes with one side much longer
  - Gaps between main data and extreme values
- **Multidimensional methods**:
  - k-NN distance
  - Isolation Forest
  - Clustering methods
  - Self-Organizing Maps
  - Dimensionality reduction tools

5.3 Treatment Approaches

5.3.1 Capping (Winsorizing)

Instead of removing outliers, put a "ceiling" and "floor". Example: If most salaries are $30k-$100k, but one is $1 million, cap it at $200k.

5.3.2 Transformation Methods

- **Log transformation** - best for income, prices, population counts
- **Square root transformation**
- Pros: Preserves all data points, improves model performance
- Cons: Reduces interpretability, requires reverse transformation

6 Discretization

6.1 Overview

Discretization divides the range of a continuous variable into intervals. Also called "binning," it can be:

- **Unsupervised** - no target variable considered
- **Supervised** - uses target variable information

6.2 Unsupervised Methods

6.2.1 Equal-Width Binning

- Divides range into n intervals of equal size
- Width: $w = (B - A)/N$ where A and B are min and max values
- Simple and fast but sensitive to outliers

6.2.2 Equal-Depth Binning

- Divides range into n intervals with approximately equal samples
- Handles outliers better, more intuitive bins
- May break natural clusters

6.2.3 Choosing Number of Bins

- **Square Root Rule**: $\sqrt{n}$ where n is number of data points
- **Rice Rule**: $2 \times n^{1/3}$
- **Business Common Sense**: Age groups, income brackets, etc.

6.3 Supervised Methods

6.3.1 Entropy-Based Discretization

Uses information theory to find optimal splits that minimize entropy:

$$Ent(S) = -\sum_{i=1}^{\#C} p_i \log_2(p_i)$$

$$Gain(S, A) = Ent(S) - \sum_{\nu \in Values(A)} \frac{\#S_\nu}{\#S} Ent(S_\nu)$$

7 Encoding

Converting categorical data into numerical format that machine learning algorithms can understand.

7.1 One-Hot Encoding

Creating binary (0/1) columns for each category:

Color_Red	Color_Blue	Color_Green
1	0	0
0	1	0
0	0	1

8 Imbalanced Learning

8.1 The Problem

Imbalanced learning occurs when there's significant asymmetry between class instances. The dominant class is the majority class, while others are minority classes.

8.2 Challenges

- Standard methods induce bias toward majority class
- Minority classes contribute less to accuracy maximization
- Misclassification costs are often non-uniform

8.3 Solutions

8.3.1 Random Undersampling

Reducing majority class instances randomly.

8.3.2 Random Oversampling

Increasing minority class instances by duplication.

8.3.3 SMOTE (Synthetic Minority Over-sampling Technique)

Generating synthetic minority class instances:

$$x_{gen} = x + \alpha \cdot (x' - x)$$

where x is a minority instance and x' is one of its k-nearest neighbors.

9 Quick Quiz

1. **Is it possible to have a useless classifier with 99% accuracy? Yes** - when the minority class is 1%
2. **Is it possible to achieve 99.9% accuracy with a trivial classifier? Yes** - when the minority class is 0.1%

10 Real-World Examples

- Credit card frauds: 2% per year
- HIV prevalence in USA: 0.4%
- Disk drive failures: 1% per year
- Factory production defects: 0.1%
- Business churn: 3%

11 Conclusion

Data preparation is a critical step in the data mining process that significantly impacts model performance. Proper handling of missing values, outliers, discretization, encoding, and class imbalance ensures that models can extract meaningful patterns from data and make accurate predictions.