CSCI316: Week 2 - Data Pre-processing Comprehensive Exam Notes

---

📖Overview

Data Pre-processing is a vital step in the data mining and machine learning pipeline. It transforms raw data into a clean and structured format suitable for analysis.

🧱Why Pre-process?

- Raw Big Data is often messy: noisy, inconsistent, incomplete.
- Mining algorithms require well-structured, high-quality data.
- Improves **accuracy, reliability**, and **efficiency** of models.

📊What is Data?

- **Data object**: Also called instance, sample, entity, etc.
- **Attributes**: Properties describing objects.
- **Dimension**: Number of attributes in a data object.

Attribute Types:

- **Numeric**: Discrete, Continuous, Fractional
- **Symbolic**: Categorical, Textual
- **Single/Multi-valued**, Compound

🛠️Attribute Conversion

Convert complex attributes into simpler, structured forms (e.g., one-hot encoding symbolic values).

🧴 Sparsity & Curse of Dimensionality

- Sparse data: Most values are zero or null.
- High dimensionality leads to inefficiency and poor model performance.

🧰Domain Understanding

- Understand **data context**: how it was collected, what it represents.
- Source reliability, attribute meaning.

📊Data Exploration

Helps select preprocessing tools and mining algorithms. - Look for: Imbalance, skew, noise, outliers, correlations. - Use **visual tools**: Histograms, Box plots, Scatter plots

---

💡Data Quality Problems

Issue	Description	Fix
Missing	Absent data points	Estimate, Predict, or Drop
Outliers	Values far from the mean	Remove, Retain, or Transform
Noise	Corrupted/inconsistent values	Filtering, Cleaning

🪜Imbalanced Data

- **Class Imbalance**: Unequal target class sizes
- **Feature Imbalance**: Skew in input features

Remedies: - **Oversampling** (e.g., SMOTE) - **Undersampling** - **Cost-sensitive learning**

📊Statistics to Measure Imbalance

- Discrete: Mode, Frequency
- Continuous: Mean, Median, Std Dev, Range, Skewness

⚖️Correlation

- Measures how two attributes vary together
- Corr ≈ 1: Strong positive, Corr ≈ -1: Strong negative

📂Data Integration

Combining multiple datasets into a unified view - Challenges: Schema mismatch, noise introduction, format inconsistency - Approaches: Schema matching, metadata analysis, domain knowledge

---

🪭Data Aggregation

- Combines records or attributes to reduce data size and variability.
- Risk: May lose useful information

📁Instance Selection & Generation

- **Selection**: Keep only useful instances (e.g., Grid Method)
- **Generation**: Create artificial data (e.g., k-means prototypes)

💊Data/Feature Transformation

- Modify attributes for better analysis
- Examples: Encoding, Normalization, Scaling, Binning

One-Hot Encoding

- Encodes categories as binary vectors
- Maintains equidistance in Euclidean space

Discretization / Binning

- Convert continuous to categorical by splitting into intervals

Normalization & Scaling

- **Min-Max**: [0,1] range
- **Z-score**: Mean=0, Std=1
- **Log-transform**: Reduce skew

📗Filtering

- Remove unwanted instances/attributes/values
- Types: Instance filters, Attribute filters, Value filters

🕑Sampling

- **Why**: Efficient when data is too large
- **Types**:
- Simple Random
- With/Without Replacement
- Stratified (preserves class ratio)

🔍Feature Generation

- Create new features to represent data more effectively
- From raw data (e.g., image edges)

- Combine existing features (e.g., density = mass/volume)

🕐Feature Selection

- Remove redundant/irrelevant attributes
- Goal: Reduce overfitting, improve efficiency
- Techniques: Correlation, Info Gain (later lecture)

🔗Summary

- Pre-processing **improves data quality**, makes mining more effective.
- **Must be tailored** to the data, problem, and mining algorithm.
- Often the most time-consuming part of the pipeline.

Would you like practice questions, diagrams, or a cheatsheet next?