Week 10: Data Transformation and Data Integration

In this preprocessing step, the data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

Data Transformation Strategies Overview

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following:

1. **Smoothing,** which works to remove noise from the data. Techniques include binning, regression, and clustering.
2. **Attribute construction (or feature construction),** where new attributes are constructed and added from the given set of attributes to help the mining process.
3. **Aggregation,** where summary or aggregation operations are applied to the data.
4. **Normalization,** where the attribute data are scaled so as to fall within a smaller range, such as -1.0 to 1.0, or 0.0 to 1.0.
5. **Discretization,** where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute.

Data Transformation by Normalization

- The measurement unit used can affect the data analysis. For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to very different results.
- In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect or “weight.”
- To help avoid dependence on the choice of measurement units, the data should be normalized or standardized. This involves transforming the data to fall within a smaller or common range such as [-1, 1] or [0.0, 1.0].
- The terms standardize and normalize are used interchangeably in data preprocessing.
- Normalizing the data attempts to give all attributes an equal weight.

- Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbor classification and clustering.
- There are many methods for data normalization. We study **min-max normalization, z-score normalization,** and **normalization by decimal scaling.**
- **Min-max normalization**
  - performs a linear transformation on the original data.

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \qquad (3.8)$$

  - **Example 1**

> **Min-max normalization.** Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$. ■

■

  - **Example 2**

Index	Value
1	100
2	300
3	250
4	150
5	500
6	400

  - For the given data, apply **min-max normalization** so that the data become in range from (-1.0 to 1.0)

- **z-score normalization**
  - In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A.

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \qquad (3.9)$$

  - **Example 1**

z-score normalization. Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600 - 54,000}{16,000} = 1.225$. ■

  - **Example 2**
    - For the previously provided data, apply z-score normalization.
  - A variation of this z-score normalization replaces the standard deviation of Eq. (3.9) by **the mean absolute deviation of A.**
  - The mean absolute deviation of A, denoted sA, is

$$s_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|). \qquad (3.10)$$

  - Thus, z-score normalization using the mean absolute deviation is

$$v_i' = \frac{v_i - \bar{A}}{s_A}. \qquad (3.11)$$

- **Normalization by decimal scaling**
  - normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A.

$$v_i' = \frac{v_i}{10^j}, \qquad (3.12)$$

  - **Example**
  -

Data Integration

Data mining often requires data integration—the merging of data from multiple data stores. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent data mining process.

- **Redundancy and Correlation Analysis**
  - **Redundancy** is another important issue in data integration. An attribute (such as annual revenue, for instance) may be redundant if it can be “derived” from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.
  - Some redundancies can be detected by **correlation analysis.** Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.
  - For **nominal data,** we use the X2 (chi-square) test.
  - For numeric attributes, we can use the correlation coefficient and covariance, both of which access how one attribute’s values vary from those of another.