Data Mining

2.5 Data Transformation

Fall 2008

Instructor: Dr. Masoud Yaghini

Data Transformation

- In data transformation, the data are transformed into forms appropriate for mining.
- Data transformation tasks:
  - Normalization: scaled to fall within a small, specified range
    - min-max normalization
    - z-score normalization
  - Attribute construction (feature construction): New attributes constructed from the given ones

Outline

- Normalization
- Attribute Construction
- References

Normalization

Normalization

- An attribute is normalized by scaling its values so that they fall within a small specified range, such as 0.0 to 1.0.
- Normalization is particularly useful for classification algorithms involving
  - neural networks
  - distance measurements such as nearest-neighbor classification and clustering.
- If using the neural network backpropagation algorithm for classification mining, normalizing the input values for each attribute measured in the training instances will help speed up the learning phase.

Normalization

- For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., ***income***) from out-weighing attributes with initially smaller ranges (e.g., binary attributes).
- Two methods for data normalization
  - *min-max normalization*
  - *z-score normalization*

Min-max normalization

- Min-max normalization performs a linear transformation on the original data.
- Suppose that:
  - $\min_A$ and $\max_A$ are the minimum and maximum values of an attribute, A.
- Min-max normalization maps a value, v, of A to v' in the range [new_min_A, new_max_A] by computing:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Min-max normalization: Example

- Let ***income*** range \$12,000 to \$98,000 normalized to [0.0, 1.0].
- Then \$73,000 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

z-score normalization

- In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean (μ) and standard deviation (σ) of A.
- A value, v, of A is normalized to v' *by computing*

$$v' = \frac{v - \mu_A}{\sigma_A}$$

z-score normalization: Example

- Let $\mu = 54{,}000$, $\sigma = 16{,}000$, for the attribute ***income***
- With z-score normalization, a value of \$73,600 for ***income*** is transformed to:

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Normalization

- Note that normalization can change the original data quite a bit, especially the z-score method.

Attribute Construction

Attribute Construction

- Attribute construction (feature construction)
  - new attributes are constructed from the given attributes and added in order to help improve the accuracy and understanding of structure in high-dimensional data.
- Example
  - we may wish to add the attribute ***area*** based on the attributes ***height*** and ***width***.
- By attribute construction can discover missing information.

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end