An Unofficial FastODT Implementation

https://github.com/Nero-DV

June 30 2024

1 Features

- Calculate entropy of attribute values.
- Compute entropy difference when an attribute value is removed.
- Determine threshold α for entropy differences.
- Detect outliers in the dataset based on the α threshold.

2 Algorithm Logic and Steps

Let X represent the dataset, where each x_i represents an attribute value. To calculate the entropy $H(X)$ of attribute values, we use the following formula based on Shannon's entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{1}$$

where $p(x_i)$ is the probability of occurrence of x_i.

To compute the entropy difference $\Delta H(X, x_j)$ when an attribute value x_j is removed, we calculate:

$$\Delta H(X, x_j) = H(X) - H(X \setminus \{x_j\}) \tag{2}$$

The threshold α is determined based on the distribution of $\Delta H(X, x_j)$ values. The outliers are detected by comparing the entropy differences to this threshold α[1].

Steps:

1. Calculate the initial entropy $H(X)$.
2. For each x_j in X, compute $H(X \setminus \{x_j\})$ and $\Delta H(X, x_j)$.
3. Determine the threshold α from the distribution of ΔH values.
4. Any x_j with $\Delta H(X, x_j)$ greater than α is considered an outlier.

References

[1] Du, Hongwei and Ye, Qiang and Sun, Zhipeng and Liu, Chuang and Xu, Wen IEEE Transactions on Network Science and Engineering, FAST-ODT: A Lightweight Outlier Detection Scheme for Categorical Data Sets, 2021, volume 8, issue 1, pages 13-24, doi 10.1109/TNSE.2020.3022869