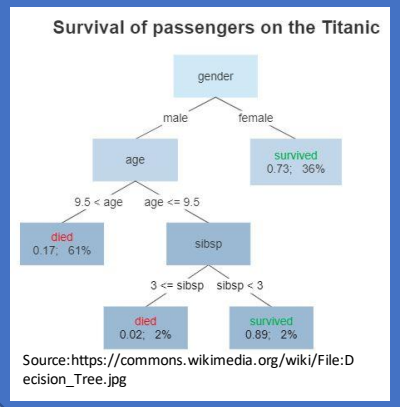
Improving Existing Optimal Decision Trees Algorithms by Redefining Their Binarization Strategy

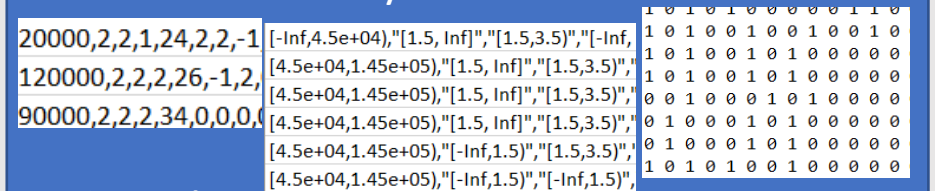
1. Introduction



Source:https://commons.wikimedia.org/wiki/File:Decision_Tree.jpg

- Decision Trees: easy to interpret data representation
- Optimal Decision Trees: best possible tree given certain size constraints, for the given dataset.
- We can't make better tree, but what if we **improve the dataset**?

2. Pre-processing data for building binary decision trees



Continuous data → Discretization → **Categorical Data** → Binarisation (encoding) → **Binary data**

How much information do we lose in data pre-processing?

3. Types of strategies

Offline: The data is processed before the algorithm launches

Online: Data is processed during the algorithm runtime

4. Discretisation strategies

Supervised:

- MDLP: minimising entropy within categories
- CAIM: maximalising interdependency with target value

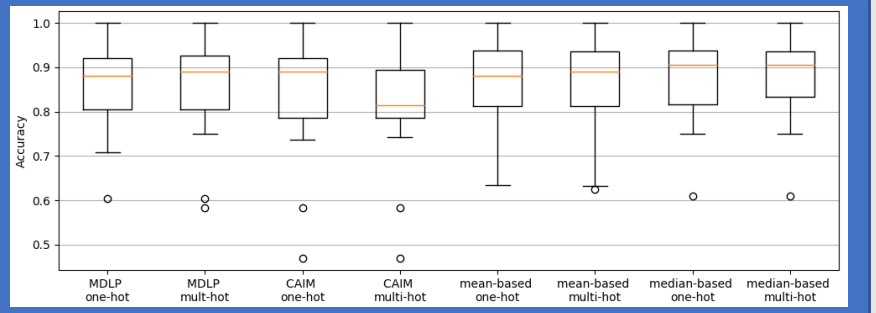
Unsupervised:

- Mean-based: attributes divided into intervals of the same size
- Median-based: attributes divided into intervals with the same amount of data instances in each

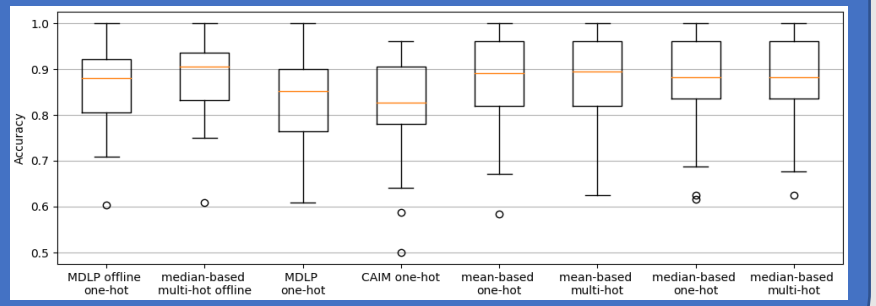
5. Encoding strategies

- One-hot encoding – each category is target value exactly once
- Multi-hot encoding – multiple combinations of categories within the scope of one attribute are target values

6. Results



Offline



Online

7. Conclusions

- Unsupervised methods, especially median-based, more successful for smaller datasets
- One-hot encoding is a useful extension that provides comparable accuracy for trees with less depth
- Online binarization quickly leads to overfitting, but improvement can be observed for datasets with high amount of data instances

TU Delft Delft University of Technology

Aleksandra Wolska

A.K.Wolska@student.tudelft.nl

Supervised by: Dr. Emir Demirović