Influence of discretization granularity on learning classification models

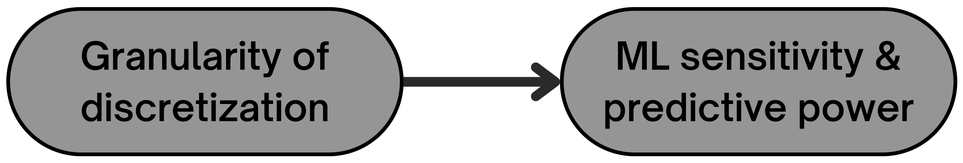
Thi Hoang Anh Tran
Malina Lara Wiesner
Dr. Ir. Maurice van Keulen

Discretization transforms continuous attributes into discrete ones by assigning them to categorical intervals. The technique has been widely used in data pre-processing for machine learning algorithms for improving model performance. Although it is well-known that discretization should maintain the distribution and patterns of the original continuous attribute, the level of granularity of the discretization method and its influence on machine learning models is largely an open question. We study this influence specifically for three classification algorithms through sensitivity analysis of parameters of five discretization methods (both supervised and unsupervised) on five datasets. This research gives insight into how discretization methods and their parameters can be combined with machine learning models to improve model performance.

1 Research Goals

RQ1: To what extent is the **discretization output sensitive to the change in discretization parameters**?

RQ2: To what extent are **machine learning models sensitive to the granularity of the discretizers** used in pre-processing?



2 Experiment design

Dataset

5 datasets from the UCI repository representing 5 different domains

Discretization

Unsupervised:
- EWD
- EFD
- FFD

Supervised:
- Chi Merge
- Decision Tree

Classification

- Categorical Naïve Bayes
- Decision Tree (ID3)
- KNN (using value difference metric)

Evaluation

Metrics:
- Inconsistency rate
- Bias & Variance
- Computation time

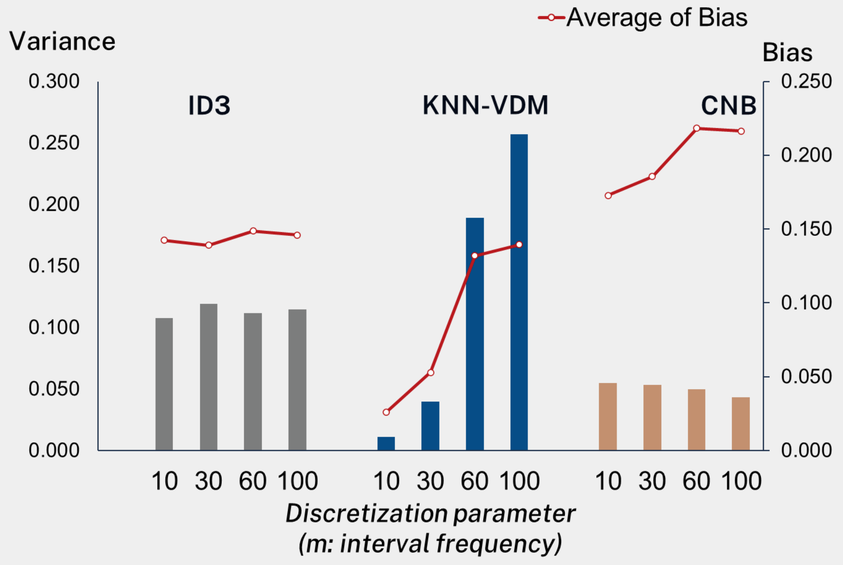
Methods:
- Sensitivity analysis
- Wilconxon signed rank test

3 Experiment results

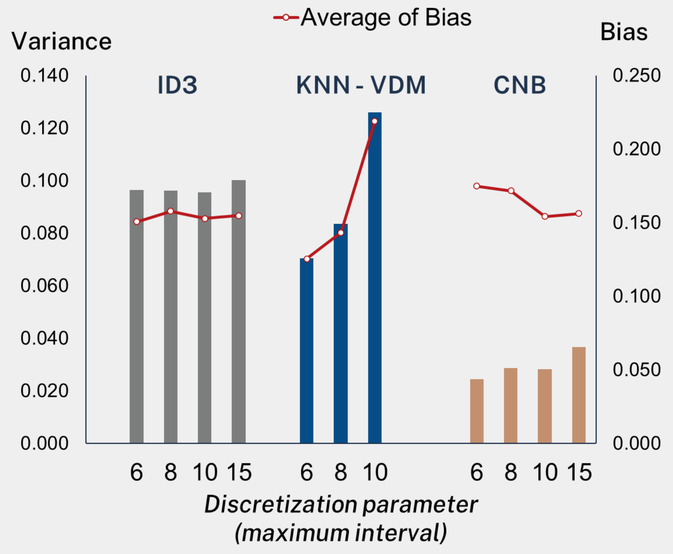
Classification Bias - Variance trade-off with increasing granularity

Increasing interval frequency (FFD) ~ Decreasing number of intervals (other discretizers)

Bias-Variance FFD

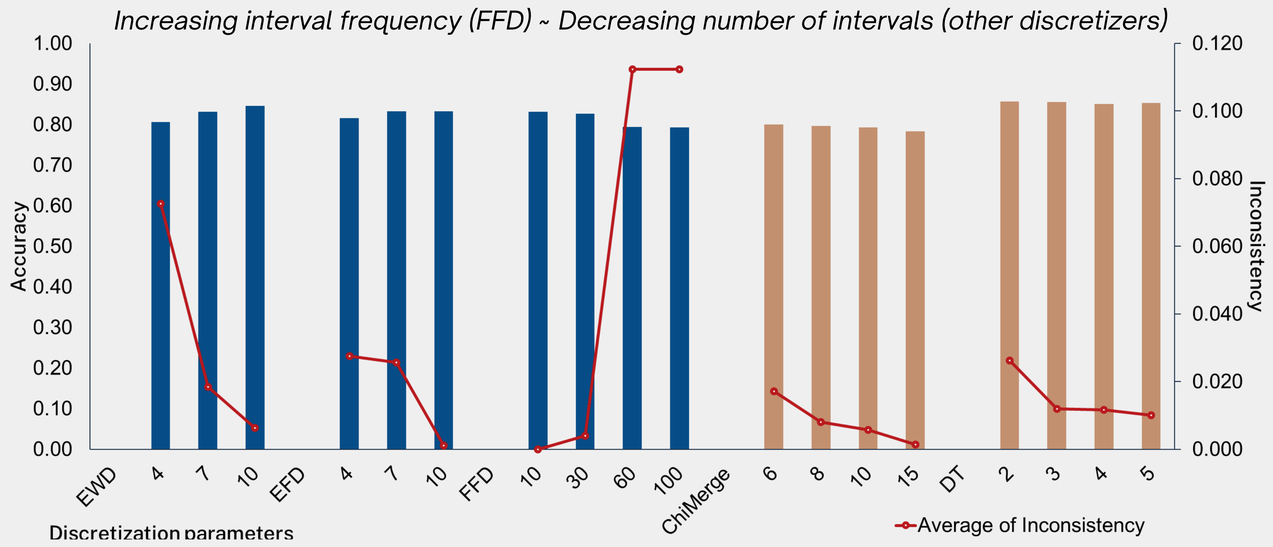


Bias-Variance ChiMerge



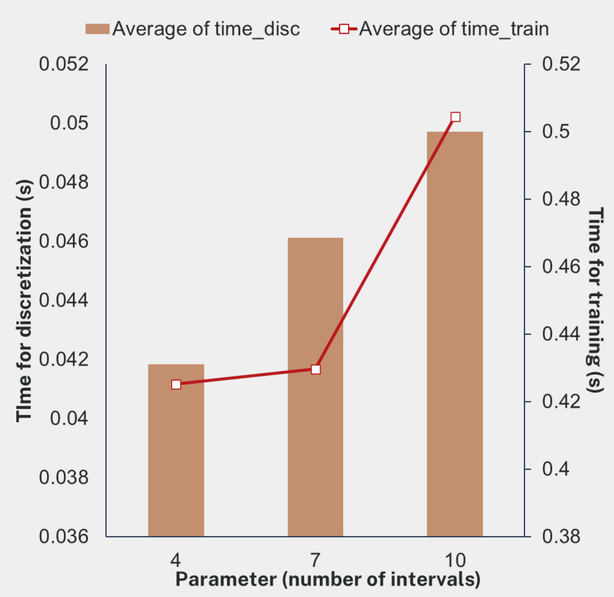
Classification accuracy - Data Inconsistency

Increasing interval frequency (FFD) ~ Decreasing number of intervals (other discretizers)

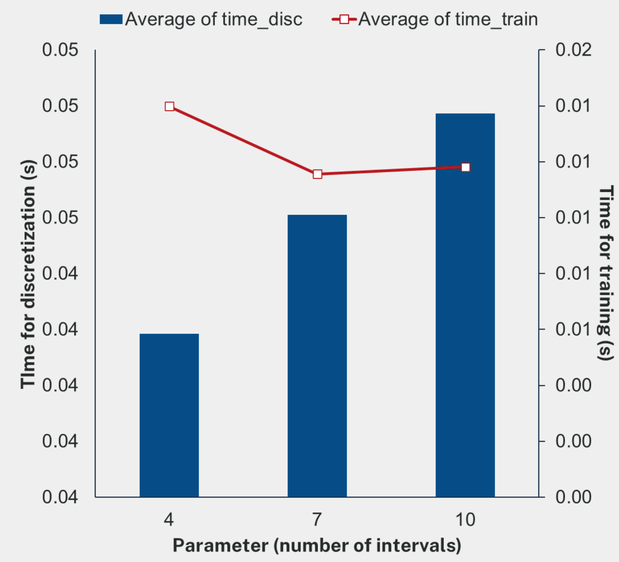


Average computation time for discretization and model training

Discretization: EWD - Model: ID3



Discretization: EWD – Model: CNB



4 Key findings

- Supervised discretization provides less flexibility in changing bin sizes and turned out less robust in implementation (**RQ1**).
- Classification models react more sensitive to unsupervised discretizers, showing higher fluctuations both in data inconsistency after discretization and classification accuracy (**RQ2**).

No perfect combination of the discretization method and ML algorithms:
- FFD (unsupervised discretizer) makes difference in classification bias
- DT (supervised discretizer) makes difference in model accuracy

We can **confirm the trade-off between discretization bias and variance** and the correlation between data inconsistency and classification accuracy