On Population Fidelity as an Estimator for the Utility of Synthetic Training Data

Alexander Florean, Jonas Forsman
({alexander.florean, jonas.forsman}@cgi.com)
CGI Karlstad
Sebastian Herold (sebastian.herold@kau.se)
Department of Mathematics and Computer Science, Karlstad University, Sweden

Objective & Research Question

In order to actually benefit from synthetic data, its utility for the intended purpose has to be ensured and, ideally, estimated before it is used to produce possibly poorly performing models. *Population fidelity (PF)* metrics are potential candidates to provide such an estimation. However, evidence of how well they estimate the utility of synthetic data is scarce.

As such, we aim to answer the following research question:

"To what degree are different population fidelity metrics capable of estimating how well ML-based classification models trained on synthetic data will perform compared to their counterparts trained on the corresponding real data?"

Experiment

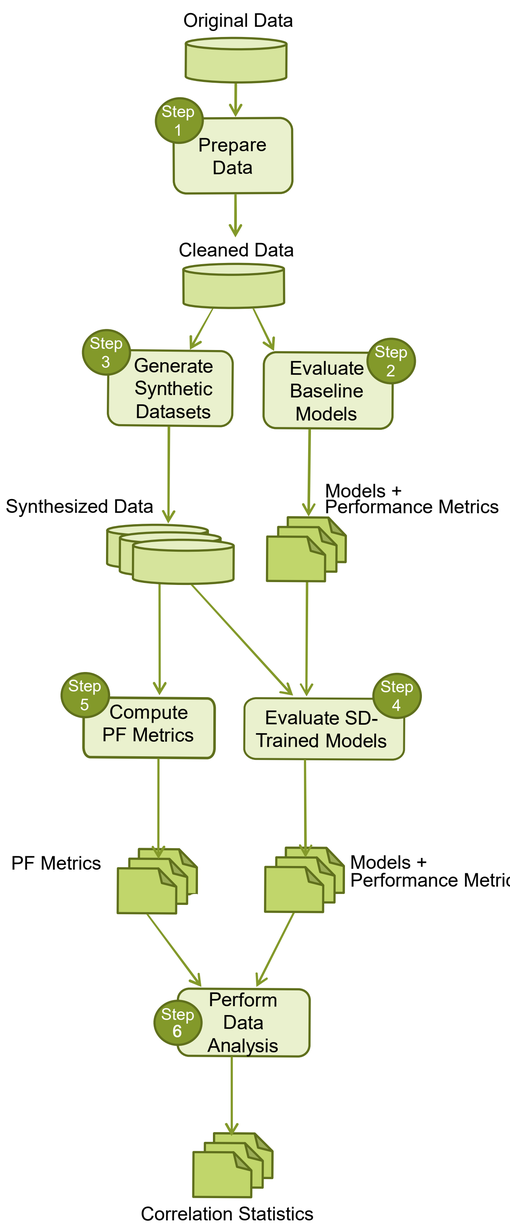
In this study, we examined

- 4 classical ML-models: K-Nearest Neighbor (KNN), Logistic Regression (LR), Random-Forest (RF), and Support Vector Machines (SVM)
- 5 publicly available and tabular datasets with independent data points
  - Adult, Bank, Diabetes, MNIST, Titanic
- 9 PF metrics (see table below)

Overview of the investigated PF metrics

Metric	Range	Value of Maximal Fidelity
BNLogLikelihood	$(-\infty, 1]$	1
Cluster Analysis	$[0, \infty)$	0
ContinuousKLD	$[0, 1]$	1
Cross Classification	$[0, 1]$	1
Chi-Statistic Test	$[0, 1]$	1
DiscreteKLD	$[0, 1]$	1
GMLogLikelihood	$(-\infty, 1]$	1
KSComplement	$[0, 1]$	1
pMSE	$[0, 0.25]$[1]	0

Experiment Design



Steps:

1. Prepare Data -
The data is cleaned and all settings for upcoming steps are defined.

2. Evaluate Baseline Models -
The models are trained, tuned and tested on the original datasets.

3. Generate Synthetic Datasets -
Synthetic datasets are synthesized with varying epochs.

4. Evaluate SD-Trained Models -
The models are trained and tuned on the synthetic datasets, then tested on original dataset.

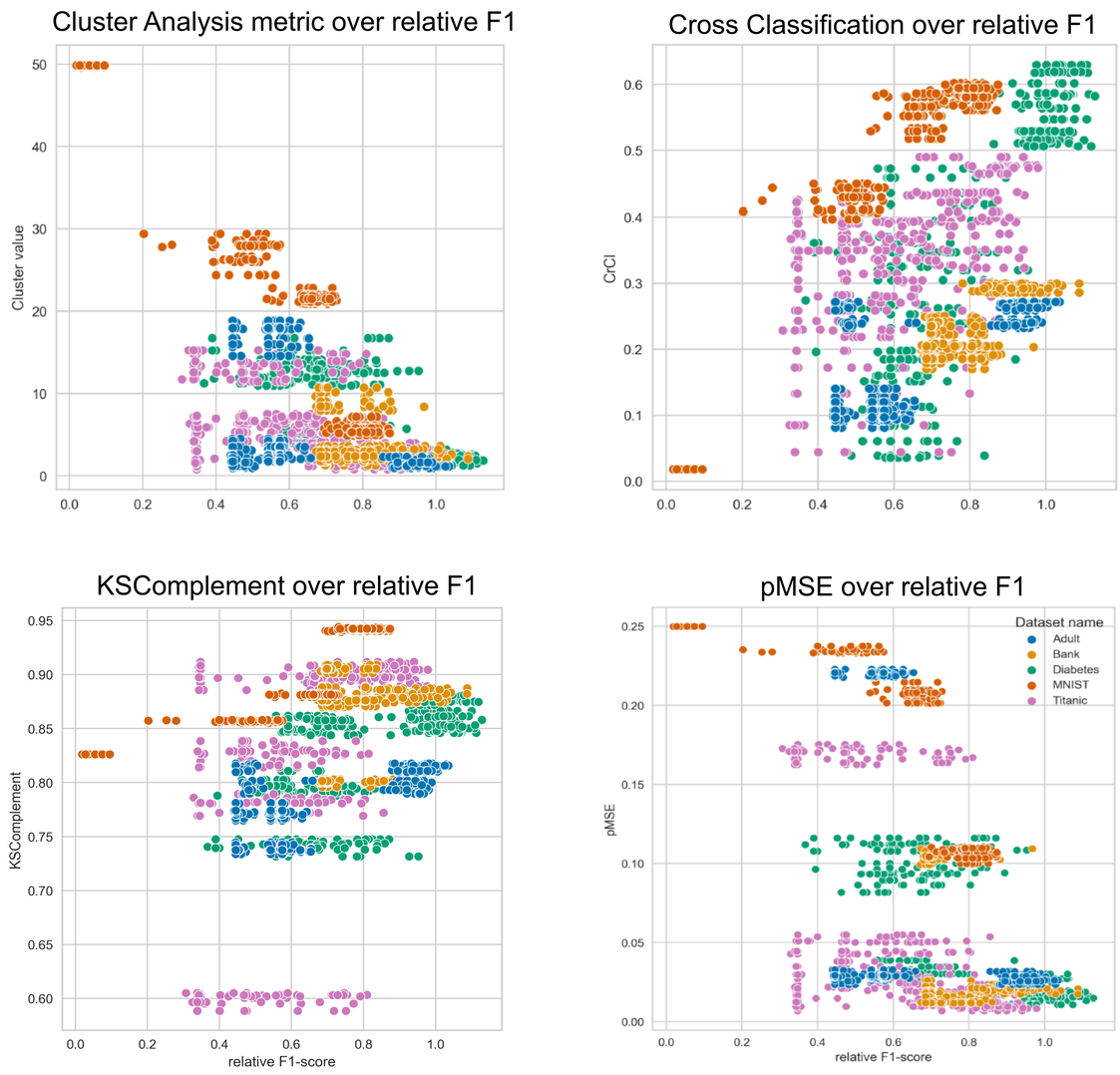
5. Compute PF Metrics

6. Perform Data Analysis -
The final analysis where the plots are created, and three statistical tests are performed.

Statistical Tests

- ***$H^A(pf)$:*** Is there a monotonic relationship between population fidelity *pf* and relative F1-score?
- ***$H^B(pf,a,t)$:*** Is there a monotonic relationship between population fidelity *pf* and relative F1-score for individual learning algorithms *a* with tuning variant *t* ?
- ***$H^C(pf,i)$:*** Is there a monotonic relationship between population fidelity *pf* and relative F1-score for individual datasets *i* ?

Results



Statistical Tests

- ***$H^A(pf)$:*** yielded statistically significant results for all metrics but GMLogLikelihood. Although, at best, the metrics showed moderate correlations.
- ***$H^B(pf,a,t)$:*** five out of the nine metrics consitently rejected the null hypothesis.
- ***$H^C(pf,i)$:*** is consistently rejected for all datasets. The Bank dataset showed weakest correlations across all metrics.

Conclusion & Future Work

Findings - Despite potential, the PF metrics examined need refinement for effective utility estimation in synthetic data applications.

Practical implications - PF metrics are currently too imprecise to infer a certain classification performance. However, they show promise in pointing out qualitative differences between synthetic datasets based on the same original dataset.

Future research direction - Investigate the influence of dataset characteristics on population fidelity measures and develop recommendations on how to measure utility in different scenarios.

Find poster at:

Read the full report at: