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

- There exists a multitude of metrics that aims to quantify synthetic data quality, out of which - *Population Fidelity (PF)* is a popular category of metrics that estimates general data similarity.
- *Utility* refers to the usefullness of synthetic data.
- Currently, evidence on how well various PF metrics ability in estimattion of synthetic data utility is scarce.

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

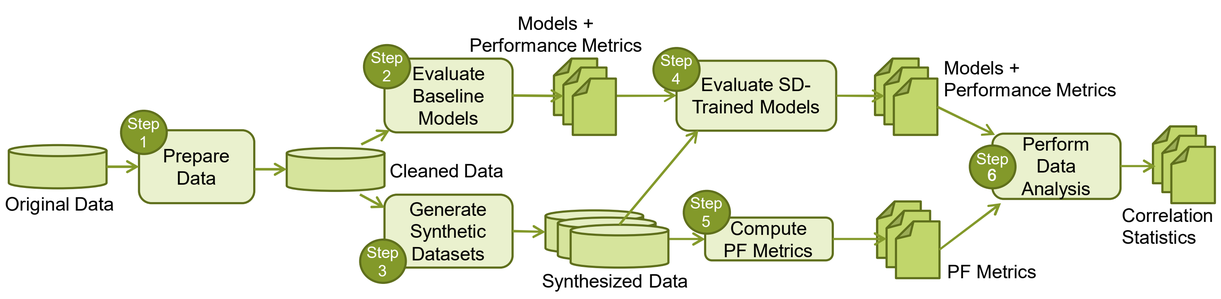
In this study, we examined:

- 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

TABLE I: Overview of the investigated population fidelity 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



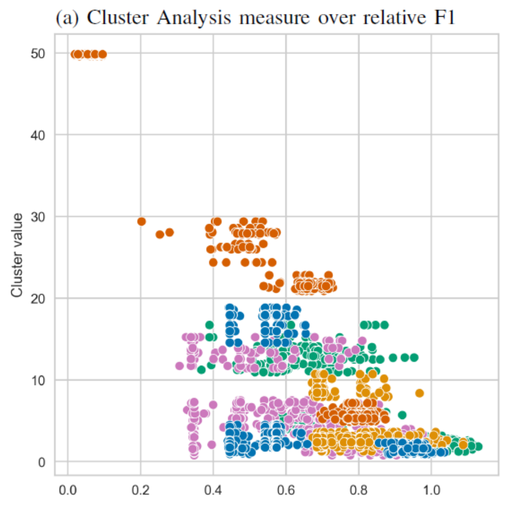
Steps:

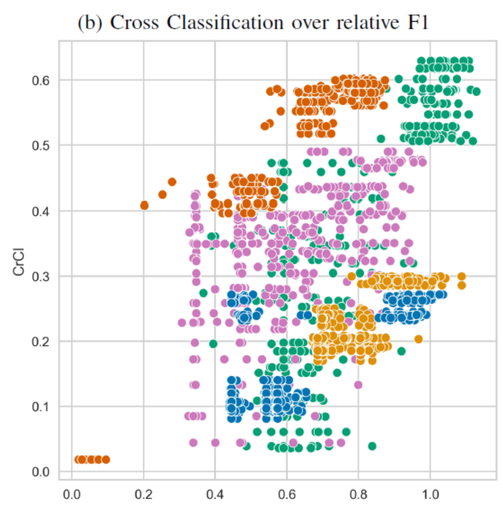
1. **Prepare Data -** The data is cleaned and all settings for upcoming steps are defined.
2. **Evaluate Baseline Models -** The models are trained, tuned and tested on the original datasets.
3. **Generate Synthetic Datasets -** Synthetic datasets are synthesized with varying epochs.
4. **Evaluate SD-Trained 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, statistical tests and plot creation is performed.

Results

$$\text{Relative f1-score}(M_{i,j}^{e,a,v}) := \frac{\text{f1}(M_{i,j}^{e,a,v})}{\text{f1}(B_i^a)}$$

, where for j=1,...,10, a synthetic dataset based on the original dataset D_i, generated by model trained for e-many epochs. Classification model $M_{i,j}^{e,a,v}$ for each synthetic dataset $S_{i,j}^e$, a referring to the algorthim used for learning and v to the tuning variant, respectively.





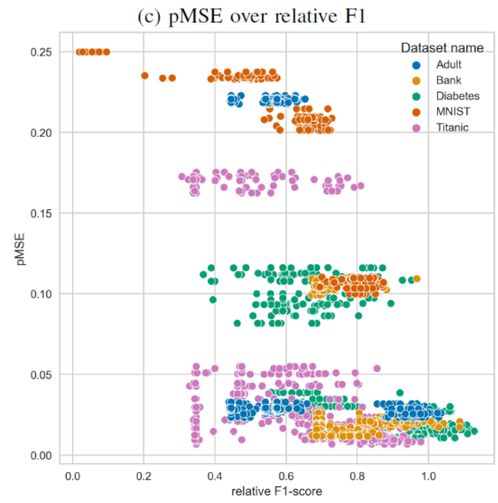


TABLE III: Results of testing $H_0^A(pf)$: Is there a monotonic relationship between population fidelity and relative F1-score?

Measure	p-value	Correlation / CI (99%)
BNLogLikelihood	0.0000	0.1761 [0.1031, 0.2471]
Cluster Measure	0.0000	-0.5370 [-0.5767, -0.4947]
ContinuousKLD	0.0000	0.2596 [0.2051, 0.3125]
CrCl	0.0000	0.4619 [0.4154, 0.506]
CSTest	0.0000	0.4300 [0.3674, 0.4887]
DiscreteKLD	0.0000	0.3414 [0.2741, 0.4055]
GMLogLikelihood	0.0188	0.0526 [-0.005, 0.1098]
KSComplement	0.0000	0.4425 [0.395, 0.4876]
pMSE	0.0000	-0.4589 [-0.5032, -0.4122]

Conclusion & Future Work

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

Practical implications - Use PF metrics as preliminary indicators rather than definitive measures.

Future research direction - Investigate how various dataset characteristics influence metric performance in utility estimation.

Read the full report at: <QR-link>