Testing Exchangeability between Real and Synthetic Data

Helena Löfström HELENA.LOFSTROM@JU.SE
Department of Computing, Jönköping University, Sweden

Lars Carlsson LARS.CARLSSON@JU.SE
Jönköping University, Jönköping, Sweden, and Centre for Reliable Machine Learning, University of London, UK

Ernst Ahlberg ERNST.AHLBERG@MOLNLYCKE.COM
Mölnlycke Health Care AB, Gothenburg, Sweden and Centre for Reliable Machine Learning, University of London, UK

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

This study introduces a method to evaluate synthetic data quality by focusing on the exchangeability of real and synthetic datasets. This is done through the use of a test martingale, which provides a statistical guarantee of the similarity of the synthetic data's representation of the original data distribution. The method was tested on six real-world datasets and their synthetic counterparts, revealing that traditional metrics such as statistical similarities and model performance may be misleading. The results indicate that the martingale test frequently rejects the hypothesis of data exchangeability, underscore the need for more robust evaluation methods. The martingale-based evaluation offers a straightforward yet effective tool to ensure that synthetic data accurately reflects the original dataset, which is essential for effective model training and validation.

Keywords: synthetic data generation, martingale testing, identical distribution, conformal transducer

1. Introduction

All predictive modelling depends on data availability, and the free flow of information is critical for efficiently utilizing AI. In practice, there are restrictions on data due to the sensitivity of the data itself. Such sensitivity may stem from intellectual property and business data such as sales. In recent years, the focus on personal data privacy has increased due to GDPR and similar legislations where the data transformations required to consider data to be anonymized make it challenging to use for the development of valuable AI models (Figueira and Vaz, 2022). The need for high-quality censored data has led to the development of methods for synthetic data generation, where data is generated based on real data (Figueira and Vaz, 2022).

The aim of these methods is to provide data that can be used interchangeably with real data to train and validate *Machine Learning* (ML) models reducing the risk of leaking sensitive data and obtain more robust and accurate models (Meister and Nguyen, 2023).

One of the core assumptions in ML is that all used data is from the same distribution or *Independent and Identically Distributed* (IID). Deviations from this assumption are called dataset shift, which could, e.g. indicate that an ML model is required to be retrained (Vovk et al., 2021). Consequently, the data must be IID, or at least exchangeable, if synthetic data is used when training or evaluating an ML model.

It is typical to evaluate models built on synthetic data with model statistics, and the underlying assumption is that when the datasets are IID, the results should be similar. For example, the authors in (Meister and Nguyen, 2023) use the *Conformal Prediction* (CP) framework to generate synthetic data and evaluate the performance of models trained on the data using the F_1 score. Similarly, in the paper presenting *Synthetic Data Vault* (SDV) (Patki et al., 2016), the authors looked at different features' distributions and the prediction accuracy of the ML models built on the synthetic data. However, the assumption that model performance tests reveal the datasets' exchangeability does not necessarily hold true.

1.1. Synthetic Data Vault

Synthetic data generation is a technique for enhancing machine learning models when real-world data is limited or restricted due to, e.g., privacy concerns. The Synthetic Data Vault (SDV) (Patki et al., 2016) is a tool that enables the creation of synthetic data that closely resembles real datasets' statistical properties and structure. The process of synthetic data generation involves building a model that captures the underlying patterns and distributions within an existing real-world dataset (El Emam et al., 2020). This model is then used to generate new, synthetic data that share the same characteristics as the original real data without compromising the privacy of the individuals or entities represented in the real dataset.

The SDV and other synthetic data generation tools employ various techniques, such as Generative Adversarial Networks (GANs) and Gaussian copulas, to ensure that the generated data closely resembles the real-world data in terms of statistical properties, relationships between variables, and overall structure (Patki et al., 2016). These approaches are effective in a wide range of applications, including healthcare, finance, and data science.

1.2. Exchangeability Martingales

A martingale results in a sequence of non-negative random variables $S_1, S_2, \ldots, S_n$, with finite expectation $\mathbb{E}[S_i] \leq \infty$ and an initial value $S_0 = 1$. In the sequence, the conditional expectation $\mathbb{E}$ of the next observation in the sequence, S_{n+1}, given all the past observations $S_1, ..., S_n$ is equal to the present value of S_n, regardless of all prior values, expressed as follows (Vovk et al., 2021, 2005):

$$\mathbb{E}[S_{n+1} \mid S_1, \ldots, S_n] = S_n \tag{1}$$

The martingale value reflects the strength of evidence against the assumption or the null hypothesis it is designed to test (Fedorova et al., 2012). If the final value of the martingale is large, it indicates a deviation from the assumption being tested, and the null hypothesis can be rejected (Vovk et al., 2021).

$$\mathbb{P}(\exists n : S_n \geq c) \leq \frac{1}{c}, \ \forall c \geq 1, \tag{2}$$

for a constant $c > 1$. This allows for the construction of *exchangeability martingales* for testing the assumption of exchangeability using conformal transducers and is applicable in areas such as anomaly detection (Vovk et al., 2005). The key to constructing an exchangeability martingale is translating the p-values into a martingale. If the distribution is exchangeable, the p-values $p_1, p_2, \dots$ output by a smoothed conformal transducer are independent and distributed uniformly on $[0, 1]$. To test the IID assumption, the exchangeability is tested by betting against the uniform distribution of the conformal p-values $(p_1, p_2, ...) \in [0, 1]^\infty$.

The remainder of this paper is structured as follows. Section 2 outlines the experimental set-ups, while the results are presented in Section 3. The paper ends (Section 4) with a discussion followed by concluding remarks.

2. Method

The aim of the proposed method is to test exchangeability between real and synthetic data and to see if such datasets can be used interchangebly in the development and application of AI models. The method can be described in four steps:

1. Data generation and preprocessing
2. Application of the conformal transducer
3. Application of the martingale test for exchangeability
4. Evaluation of the datasets

described in Subsections 2.1 through 2.4.

2.1. Data Generation and Preprocessing

This study investigates the exchangeability between real-world datasets and synthetic datasets created using the generative AI tool SDV, (Patki et al., 2016). For each experiment, a real-world dataset will be selected and used as input for SDV to generate a corresponding synthetic dataset of the same number of instances. We use the following datasets **`adult-sdv`**[1], **`credit-g`**, **`spambase`**, **`qsar-biodeg`**, **`adult`** and `RWI`[2]. More information about the datasets can be found in Table 1. The synthetic datasets are created in two different ways, but both using the software package `sdv`. The first way only applies to **`adult-sdv`** and here the function `download_demo` is used and we completely rely on the functionality of `sdv` in selecting how to treat the different types of data. For all other datasets, the function `CTGANSynthesizer` is used with the package's automatic detection of column types. It is worth noting that the **`adult`** dataset is slightly larger than the **`adult-sdv`** dataset although they represent the same underlying problem. Both the real and synthetic datasets undergo identical preprocessing steps to ensure compatibility for calculating the p-values. Non-numeric features are transformed using the `OneHotEncoder` (Pedregosa et al., 2011) column by column.

[1] The actual name of the dataset is adult, but this version comes from SDV.

[2] RWI is an acronym introduced here instead of using the original name Run_or_walk_information.

Dataset Name	Instances	Features	Numeric Features	Source
adult-sdv	32561	15	6	(Patki et al., 2016)
credit-g	1000	21	7	(Vanschoren et al., 2013)
spambase	4601	58	58	(Vanschoren et al., 2013)
qsar-biodeg	1055	42	42	(Vanschoren et al., 2013)
adult	48842	15	6	(Vanschoren et al., 2013)
RWI	88588	7	7	(Vanschoren et al., 2013)

Table 1: Additional information on the datasets used in this study.

2.2. Conformal Transducer

To assess exchangeability, we need to calculate p-values, which are defined as

$$p_i = \frac{|\{i|\alpha_i < \alpha_n\}| + \theta_n|\{\alpha_i = \alpha_n\}|}{n}, \tag{3}$$

where θ_i is drawn from the uniform distribution on $[0, 1]$ independently of everthing else and i is taken from the range $1, \ldots, n$. Furthermore, $\alpha_i = A(z_i)$, which is an inductive nonconformity score, where $A()$ is defined by the Python package `crepes` (Boström, 2022). This package utilizes a machine-learning model to generate nonconformity scores. In this study, we employ a Random Forest classifier from the `scikit-learn` library (Pedregosa et al., 2011) as the underlying machine-learning algorithm, and the corresponding `prob_a` is used to generate nonconformity scores.

2.3. Martingale Test for Exchangeability

A martingale test will be employed to statistically evaluate the exchangeability between the real and synthetic datasets based on the p-values obtained from the conformal transducer. We follow the work in (Vovk et al., 2021) and use the martingale described in Algorithm 1 for the test. Since we expect S_n to become very large for tests that fail we will work with $\mathtt{lg}(S_n)$ in the following.

Algorithm 1: Simple Jumper $(p_1, p_2, \ldots, p_n) \mapsto (S_1, S_2, \ldots, S_n)$

```
$C_{-1} := C_0 := C_1 := \frac{1}{3}$
$C := 1$
$J = 0.01$
for $k = 1, 2, \ldots, n$ do
    for $\varepsilon \in \{-1, 0, 1\}$ do
        $C_\varepsilon := (1 - J)C_\varepsilon + (J/3)C$
    end
    for $\varepsilon \in \{-1, 0, 1\}$ do
        $C_\varepsilon := C_\varepsilon(1 + \varepsilon(p_k - 0.5))$
    end
    $S_k := C := C_{-1} + C_0 + C_1$
end
```

2.4. Evaluation on Machine Learning Datasets

We will compare the real and synthetic datasets by applying the conformal transducer in four scenarios:

- Real on Real (R-R): The conformal transducer is trained and tested on the real dataset to verify that our test works.
- Synthetic on Synthetic (S-S): The conformal transducer is trained and tested on the synthetic dataset, similar to the R-R case.
- Real on Synthetic (R-S): The conformal transducer, trained and calibrated on the real dataset, generates p-values for the synthetic dataset.
- Synthetic on Real (S-R): The conformal transducer, trained and calibrated on the synthetic dataset, generates p-values for the real dataset.

In each scenario, half of the datasets will be randomly sampled as training data and the other half as test data. The training data will then be further divided into two equal parts (each $\frac{1}{4}$ of the dataset) to form proper training data and calibration data, which will be used to define the nonconformity function and provide nonconformity scores, respectively. This process will be repeated 50 times for the real and synthetic data, following the same division method used in Conformal Prediction.

3. Results

Results from the application of the method are presented in Tables 2, 3, 4 and 5; one for each scenario defined in Section 2.4. Each table contains summary statistics from the 50 runs per dataset of the log-transformed martingale values. If the mean value of all logarithmic martingale values is large, it can be considered as evidence against the hypothesis that two different datasets come from the same underlying probability distribution as stated in Equation (2).

Tables 2 and 3 show that when test and training data is drawn from the same dataset, the martingale values tend to 0, thus the data is exchangeable. When however, training and test data are not drawn from the same dataset, ie Tables 4 and 5, the martingale values grow rapidly and tend to infinity indicating that training and test data is not exchangeable.

It is notable that the *adult* passes the test in the R-S and S-R scenarios, whereas *adult-sdv* only passes the test in the S-R scenario. For all other datasets, the S-R and R-S scenarios fail.

Dataset	Mean	Std	Min	Max
adult-sdv	-63.61	6.58	-75.52	-45.10
credit-g	-1.40	1.87	-3.49	5.05
spambase	-8.48	2.56	-12.51	-1.85
qsar-biodeg	-2.01	1.84	-3.87	7.06
adult	-96.63	7.09	-109.32	-79.45
RWI	-174.65	7.36	-192.76	-159.12

Table 2: Some metrics on the distribution of the martingale values for the 50 runs in the R-R case.

Dataset	Mean	Std	Min	Max
adult-sdv	-63.34	4.93	-73.59	-48.57
credit-g	-1.81	1.82	-3.49	6.04
spambase	-8.73	2.49	-12.45	-1.68
qsar-biodeg	-1.45	3.04	-3.73	16.17
adult	-96.58	5.53	-106.71	-82.01
RWI	-174.72	10.01	-195.17	-156.26

Table 3: Some metrics on the distribution of the martingale values for the 50 runs in the S-S case.

Dataset	Mean	Std	Min	Max
adult-sdv	76.50	35.83	8.83	169.04
credit-g	57.72	8.62	36.61	72.18
spambase	519.12	26.78	440.35	579.82
qsar-biodeg	127.23	6.70	111.12	140.12
adult	-8.35	26.97	-62.07	60.97
RWI	709.78	0.00	709.78	709.78

Table 4: Some metrics on the distribution of the martingale values for the 50 runs in the R-S case.

Dataset	Mean	Std	Min	Max
adult-sdv	-65.53	5.26	-73.60	-52.25
credit-g	2.93	6.28	-3.25	26.78
spambase	29.82	14.30	1.04	57.76
qsar-biodeg	13.87	13.68	-3.46	57.60
adult	-79.51	15.20	-99.85	-37.54
RWI	708.55	5.21	684.25	709.78

Table 5: Some metrics on the distribution of the martingale values for the 50 runs in the S-R case.

4. Discussion and Conclusion

This methodology provides a framework for evaluating the fidelity of synthetic datasets in replicating the exchangeability property of their real counterparts. This study contributes to the ongoing discourse on the utility and limitations of generative AI models in creating reliable synthetic data for diverse applications. In particular, it illustrates the limitations of synthetic data usage for predictive modeling under the exchangeability assumption. This is the case when e.g. building predictive models using Conformal Prediction, since datasets produced by GAN-methods do not guarantee valid predictions unless they were specifically designed for this purpose.

Future work could include the extension of the current method to regression datasets and the generative algorithm described in (Meister and Nguyen, 2023). Another aspect to investigate is the distribution of martingales for repeated experiments, looking into the high maximum martingales for R-R and S-S.

Acknowledgments

The authors acknowledge the Swedish Knowledge Foundation and industrial partners for financially supporting the research and education environment on Knowledge Intensive Product Realisation SPARK at Jönköping University, Sweden. Project: ETIAI grant no. 20230040.

References

Henrik Boström. crepes: a python package for generating conformal regressors and predictive systems. In Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo, and Lars Carlsson, editors, *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction and Applications*, volume 179 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. *arXiv preprint arXiv:1204.3251*, 2012.

Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and gans. *Mathematics*, 10(15):2733, 2022.

Julia A. Meister and Khuong An Nguyen. Conformalised data synthesis with statistical quality guarantees, 2023.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016. doi: 10.1109/DSAA.2016.49.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning, 2013. URL https://www.openml.org. Accessed: April 11, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications*, pages 191–210. PMLR, 2021.