Data Synthesis for Fairness Audits of Learning Analytics Algorithms

Linda Fernsel [1] and Katharina Simbeck [2]

Abstract: The purpose of methods of fairness auditing is to uncover to what extent Learning Analytics algorithms are fair. Fairness auditing methods often rely on pre-existing test data. In the context of Learning Analytics auditing, learning data is needed for testing. However, learning data might not be available (in large quantities) due to privacy concerns. Our poster shares our findings on how relational data for fairness audits of Learning Analytics systems can be synthesized from little pre-existing data, using the most promising available data synthesizers.

Keywords: Learning analytics, synthetic data, fairness audit

1 Introduction

Learning Analytics is the field concerned with making sense of learning data, enabling stakeholders in learning processes (students, teachers, administrators) to better understand a learning process or environment and possibly adapt it to individual learners (cf. [La22] (p. 8ff)). Learning management systems like Moodle can integrate Learning Analytics, connecting "delivery of educational materials" to "student activity" [La22](p. 10). For instance, Moodle tries to predict which students might fail to complete a course [Mo18]. By conducting fairness audits one can find out whether Learning Analytics algorithms, like Moodle's dropout prediction model, discriminate against groups of learners (see for example [RS19]).

Data-based fairness audits require test data. The challenge is that test data is not always available, especially not in the context of Learning Analytics [Be16][Do19]: Learning data contains sensitive user data generated over long periods. Therefore the need to synthesize realistic learning data arises.

In section 2 we give a brief overview of related work and highlight how we will contribute to the existing research body. In section 3 we outline our approach. We end with a summary of the content of this poster submission in section 4.

1 University of Applied Sciences (HTW), 10313 Berlin, fernsel@htw-berlin.de, https://orcid.org/**0000-0002-0239-8951**
2 University of Applied Sciences (HTW), 10313 Berlin, simbeck@htw-berlin.de, https://orcid.org/0000-0001-6792-461X

https://doi.org/10.30844/AKWI_2022_21

2 Related Work

The challenge of synthesizing data can be approached via statistical sampling methods (statistical methods) or methods from the field of machine learning (neural methods) [Fa20]. Examples of statistical methods are "copulas, Bayesian Networks, Gibbs sampling and Fourier decomposition" [Fa20]. Variational auto-encoders and generative adversarial networks are examples of neural methods [Fa20]. There are various tools available that implement statistical and/or neural data synthesis methods for realistic tabular data, "AIT-EST" [SAH22] or the "Synthetic Data Vault" for instance [Pa16].

Learning data have been synthesized on several occasions. [Wu15] devises a method, Latent-Explicit Markov Model, for synthesizing realistic MOOCdb learning activity data that can be used in Learning Analytics model training. "DATASIM" is a software for generating "xAPI" (a learning record standard) data, parameterized with information on potential users and user groups, with the goal of testing systems using xAPI [Bl20]. [WLS13] use synthetic learning data for machine learning model performance testing. [Vo21] evaluate one neural- and one statistical method's usefulness for the generation of privacy-preserving learning data and find that the neural method creates more useful results.

Even though advances have been made in the field of data synthesis, the generation of realistic tabular remains challenging, e.g. because columns have different data types or unknown distributions [Fa20], [Xu19]. None of the named methods and tools can synthesize Moodle learning data in an out-of-the-box way.

2.1 Our Contribution

The topic of our poster is how to synthesize realistic, tabular, relational data that can be used for the fairness audit of a machine learning model in the field of Learning Analytics.

Our work focuses on applying available method(s) of data synthesis for generating realistic, tabular learning data for fairness audits of Moodle's classification model of students at risk of dropping out. We will evaluate whether different learning styles are represented in the synthetic data and whether the synthetic data can be used for auditing the dropout prediction model.

As synthetic data is necessary for a variety of goals, including but not limited to machine learning model training [Fa20][Bo21][Ur21], machine learning model evaluation [WLS13][Be16][SAH22], and software testing [LMG14][Be16][EMH20], our findings apply in general to the use of synthetic data in machine learning, quality assurance or in learning analytics.

3 Approach

To synthesize learning data specific to the Moodle Learning Management System, we follow the general approach for data synthesis suggested in [EMH20], combined with suggestions from [Fa20] and [SAH22]. [EMH20] break down a general data synthesis process into three basic steps. First, if available, real data or data from "existing models or simulations" is prepared [EMH20]. Next, synthetic data is generated. In the third and last step, the synthesized data's utility is assessed [EMH20].

The data preparation step can include removing data errors, making data consistent, harmonization, linking data [EMH20], encoding and normalization of pre-existing or simulated data values, and transforming the data into the necessary shape [Fa20].

The data generation step concretely consists of three parts: (1) formulating data requirements in the form of column and association constraints [SAH22], (2) creating data, e.g. by training a model with real data and then generating data with the trained model, and (3) post-processing created data [Fa20].

We select the most promising available solutions for synthesizing tabular data and try them in the use case of generating data for Moodle's Learning Analytics algorithm for dropout prediction. Utility assessment (cf. [EMH20]) allows evaluating how useful each tried data synthesis method is.

Challenges are (1) that only little pre-existing data is available for seeding and utility assessment, (2) that data needs to realistically represent different learning styles, and (3) that created data needs to be integrated into the Moodle Learning Management System to be used in the fairness audit of Moodle's dropout prediction model.

4 Summary

We share experiences and findings from following a three-step data synthesis approach - seed data preparation, data generation, and utility assessment – for creating realistic, tabular learning data.

5 Acknowledgments

This research is funded by the Federal Ministry of Education and Research of Germany (project nr: 16DHB4002). Thanks to André Selmanagic and three anonymous reviewers for their feedback.

Literature

[Be16] Berg, A. M. et. al.: The Role of a Reference Synthetic Data Generator within the Field of Learning Analytics. Journal of Learning Analytics, Vol. 3 (1), P. 107-128, 2016.

[Bl20] Blake-Plock, S: DATASIM – Data and Training Analytics Simulated Input Modeler. Technical Report, Yet Analytics, Baltimore, Maryland, USA, 2020.

[Bo21] Bourou, S. et. al.; A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. Information, Vol 12 (9), 2021.

[Do19] Dorodchi, M. et. al.: Using Synthetic Data Generators to Promote Open Science in Higher Education Learning Analytics. In (IEEE): International Conference on Big Data (Big Data), Los Angeles, 2019. P. 4672-4675.

[EMH20] El Emam, K.; Mosquera, L.; Hoptroff, R.: Practical Synthetic Data Generation, O'Reilly Media, Inc., Sebastopol, USA, 2020.

[Fa20] Fan, J. et. al: Relational Data Synthesis Using Generative Adversarial Networks: A Design Space Exploration. Proceedings of the VLDB Endowment, 2020. Vol. 13 (12), P. 1962-1975.

[La22] Lang, C. et. al. (Eds): Handbook of Learning Analytics (2nd ed.). Society of Learning Analytics Research (SoLAR), Vancouver, 2022.

[LMG14] Lu, W.; Miklau, G.; Gupta, V.: Generating private synthetic databases for untrusted system evaluation. In (IEEE): 30h International Conference on Data Engineering, Chicago, 2014. P. 652-663.

[Mo18] Monllaó Olivé, D. et. al.: A supervised learning framework for learning management systems. In: Proceedings of the First International Conference on Data Science, E-learning and Information Systems (DATA), Madrid, 2018.

[Pa16] Patki, N.; Wedge, R.; Veeramachaneni, K.: The Synthetic Data Vault. In (IEEE): International Conference on Data Science and Advanced Analytics (DSAA), Montreal 2016. P. 399-410.

[RS19] Riazy, S.; Simbeck, K.: Predictive Algorithms in Learning Analytics and their Fairness. In (Gesellschaft für Informatik e.V. (GI)): DeLFI – Die 17. Fachtagung Bildungstechnologien, Berlin, 2019. P. 223-228.

[SAH22] Saha, D.; Aggarwal, A.; Hans, S.: Data Synthesis for Testing Black-Box Machine Learning Models. In (ACM): 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), Bangalore, India, 2022. P. 110-114.

[Ur21] Urbina, F. et. al.: MegaSyn: Integrating Generative Molecule Design, Automated Analog Designer and Synthetic Viability Prediction. ACS Omega, Vol 7 (22), P. 18699-18713, 2022.

[WLS13] Waters, A. E.; Lan, A. S.; Studer, C.: Sparse probit factor analysis for learning analytics. In (IEEE): International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013. P. 8776-8780.

[Wu15] Wu, M. Q.: The synthetic student: a machine learning model to simulate MOOC data. Master thesis, MIT, Department of Electrical Engineering and Computer Science, Cambridge, 2015.

[Vo21] Vostres, Y.: Generierung synthetischer Daten – Betrachtung hinsichtlich Datenschutz, Innovationspotential und technischer Umsetzung. Bachelor thesis, University of Applied Sciences (HTW), Berlin, 2021.

[Xu19] Xu, L. et. al: Modeling Tabular data using Conditional GAN. In (Curran Associates, Inc.): Advances in Neural Information Processing Systems 32 (NeurIPS), Vancouver, 2019. P. 7335-7345.