Using Synthetic Data Generators to Promote Open Science in Higher Education Learning Analytics

Mohsen Dorodchi*, Erfan Al-Hossami*, Aileen Benedict * and Elise Demeter†
*Department of Computer Science
University of North Carolina, Charlotte
Charlotte, NC
Email: (Mohsen.Dorodchi, ealhossa, abenedi3)@uncc.edu
†Office of Assessment and Accreditation
University of North Carolina at Charlotte
Charlotte, NC
Email: edemeter@uncc.edu

***Abstract*—Data sharing is a common contribution to open science. The creation of open datasets can speed up research advancements by allowing researchers to focus efforts on developing and validating analytical techniques, rather than on obtaining data. Open datasets also allow researchers to benchmark new analytical approaches against a known standard, and increase the reproducibility of research. The field of higher education learning analytics could benefit from the creation of open, shared datasets on higher education students as these data do not currently exist in open and accessible formats. Here, we propose the use of synthetic data generators to create open access versions of student data. Synthetic datasets have an advantage over real data, as private student data is protected by federal laws. We compare the characteristics of the synthetic data to the original data and illustrate a model for how the synthetic data can be leveraged for developing and optimizing a common learning analytics algorithm.**

***Index Terms*—learning analytics, collaborative research, open science**

I. Introduction

Open science is the movement to make scientific research, including data, publications, and methodologies, accessible to everyone [1]. According to a review done by Vicente-Sáez and Martínez-Fuentes, open science can be defined as "transparent and accessible knowledge that is shared and developed through collaborative network" [2]. This is becoming increasingly important for scientific research, as its shared transparency can help accelerate research and will have an effect on the overall process [1], [3]. According to a recent article in Nature, the trend in open science is "moving towards a greater openness, in terms of not just data but also "publications", computer code, and workflows" [4].

Learning analytics is an emerging field and has been growing as more practitioners, institutions, and researchers see its potential [5]. The Society for Open Learning Analytics Research (SoLAR) has defined it as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [6]. SoLAR has also been contributing towards this field "to guide learners, educators, administrators, and funders in making learning-related decisions" [7]. Researchers in this field have distinguished several methods used to examine learners' data and have categorized them as (1) data mining techniques; (2) statistics and mathematics; (3) text mining, semantics, and linguistic analysis; (4) visualization; (5) social network analysis; (5) qualitative analysis; and (6) gamification [6]. Of these categories, it was found that both data mining techniques (such as classification, clustering, and predictive analytics) and studies involving statistics and mathematics for data analysis were the two most employed methods [6]. In another paper, the authors state that "one of the main applications of learning analytics is tracking and predicting learners' performance as well as identifying potential problematic issues and students at work" [8]. All of this further emphasizes the need for data for the learning analytics to thrive.

However, there are challenges with data tracking, data collection, and data analysis, as well as ethical concerns with legal and privacy issues [8], [9]. For example, many things must be considered during the process of collecting data, such as the availability of resources and potential gaps due to "the inability to share proprietary information gathered by the institution" [8]. Because of the sensitive nature of data, "formal data-sharing agreements are too slow and expensive to create in ad hoc situations" [10]. These are all challenges that make open data, one aspect of open science, extremely difficult, especially in the field of learning analytics.

This difficulty in having open data for learning analytics can then lead to further challenges. The nature of learning analytics projects is often exploratory due to data complexity, sparsity, and heterogeneity. As mentioned, the sharing of possible solutions and data is not a common practice. Therefore, collaboration amongst researchers is not a straightforward process. For example, starting a project can be difficult as gaining access to a real dataset may take a long time. An open dataset would allow researchers to begin work on other aspects of a project in parallel to the data collection process.

However, there seems to be a lack of standard baseline datasets and solution models for common learning analytics problems.

We believe that it is important to start discussing how to facilitate the concept of open science, including open data, in the field of learning analytics so that (1) collaboration may become a more straightforward process, (2) more benchmark datasets can become available for better comparisons of various methods, and (3) the field itself can advance further through the shared transparency that the movement of open science brings. We propose that synthetic data generators can be used in this field to help solve some of the challenges mentioned regarding the data, and can help to begin creating open benchmark datasets. In the next section, we will review the concept of data generators and some existing work in this area. We will then examine the application of an existing data synthesizer to one of our university's datasets and compare the distributions of the original dataset features with that of the synthetic dataset.

II. Review of Synthetic Data Generators

It is stated that one issue in evaluating data analysis algorithms, in general, is "the availability of representative data" [11]. However, when sharing data between organizations, privacy and confidentiality is extremely important to upkeep [12]. This can create a barrier when trying to share data, especially when striving to promote open data overall. Synthetic data is oftentimes used as a solution to "avoid accidental disclosure or reconstruction of information" [13]. Thus, it may also be a solution to further promote open data.

Data synthesis has been used to create shareable data in light of privacy constraints. Ping et. al present a tool called DataSynthesizer that "takes a sensitive dataset as input and generates a structurally and statistically similar synthetic dataset with strong privacy guarantees" [10] and eliminates the dependency for the real data [14]. This is achieved using three modules: (1) the DataDescriber examines the existing data to find correlations and distributions, (2) the DataGenerator then samples from the summary provided by the DataDescriber to generate synthetic data, and finally (3) the ModelInspector shows a description allowing the data owner to assess the synthetic data. Pudjijono and Christen developed another method for generating personal information where (1) several original records are first created (using real data), and then (2) of those original records, randomly selected ones are modified, duplicated, and stored [12]. Patki et. al present the Synthetic Data Vault (SDV), another system for creating synthetic data, as well as their work with generating five different publicly available datasets [14].

Data synthesis has also been used in various specific fields. For example, Kinney et. al created a synthetic, public version of the Longitudinal Business Database (LBD) [15] to "allow researchers to access data more efficiently" [16]. Other areas include work with: weather behavior, like generating wind data [17], fraud detection [18], and generating network workload [19]. The use of synthetic data generators has also been explored in the field of learning analytics. Berg et. al mentions demand for a "tool generating a wide range of synthetic data" and that synthetic data will "accelerate the creation of complex and layered learning analytics infrastructure" [13]. There have been scattered endeavors such as the Learning Analytics Initiative[1] that have incorporated data synthesis [13]. However, despite the need and initial steps, synthetic data generation and open data are still an uncommon practices in learning analytics.

III. Application of Data Synthesis

In this section, we propose data synthesis as a solution to the problem of data shareability in the learning analytics field and demonstrate a comparison of the original and synthetic datasets. Our original dataset consisted of undergraduate records from a four-year, public university. The dataset contained over 400 features on a large sample of 7,206 undergraduate students who were admitted to the institution as first-time-in-college students between the years 2010-2013. For this experiment, we hand-selected 20 features that were most relevant to a classification task. We then generated a synthetic dataset of the same number of samples as the original (7,206 samples) and all 20 features. A review of synthetic data generator works was conducted in section II. We chose to apply DataSynthesizer [10] in its correlated attribute mode due to the correlated nature of the features in the dataset to generate a synthetic dataset resembling that of the original. Our choice was motivated by its methodology, ease of implementation, and accessibility of its code. As for the parameters, the threshold value was 20, which if the domain size of the feature is less than 20 it's considered a categorical feature. Epsilon was set to 0.1. The higher epsilon value the lower the injected noises. And lastly, we set maximum number of parents in the Bayesian network to 2. We first started with loading the the original dataset and creating a DataDescriber object and passing the threshold value to it. Then we use the DataDescriber method *describe_dataset_in_correlated_attribute_mode()* and passing it the original data, epsilon, and the maximum number of parents in the Bayesian Network. That creates the data description file, which describes each of the features' correlations and distributions. After that, we call the method *save_dataset_description_to_file()* passing it the filename $descriptionFile$. Lastly, we instantiate the DataGenerator object and call its method *generate_dataset_in_correlated_attribute_mode()* passing it $descriptionFile$, to generate the synthetic dataset from the resulting data description file. The Data synthesis approach is illustrated in Algorithm 1.

The synthetic dataset does not contain rows that are real nor identifiable, hence it would be shareable with other researchers without violating students' privacy. The dataset would still be representative of the original and valid for learning analytics purposes. To demonstrate the resemblance of the synthetic dataset, we compare the real and synthetic distributions for the following selected features: age (computed as of the

[1]https://confluence.sakaiproject.org/display/LAI/Learning+Analytics+Initiative

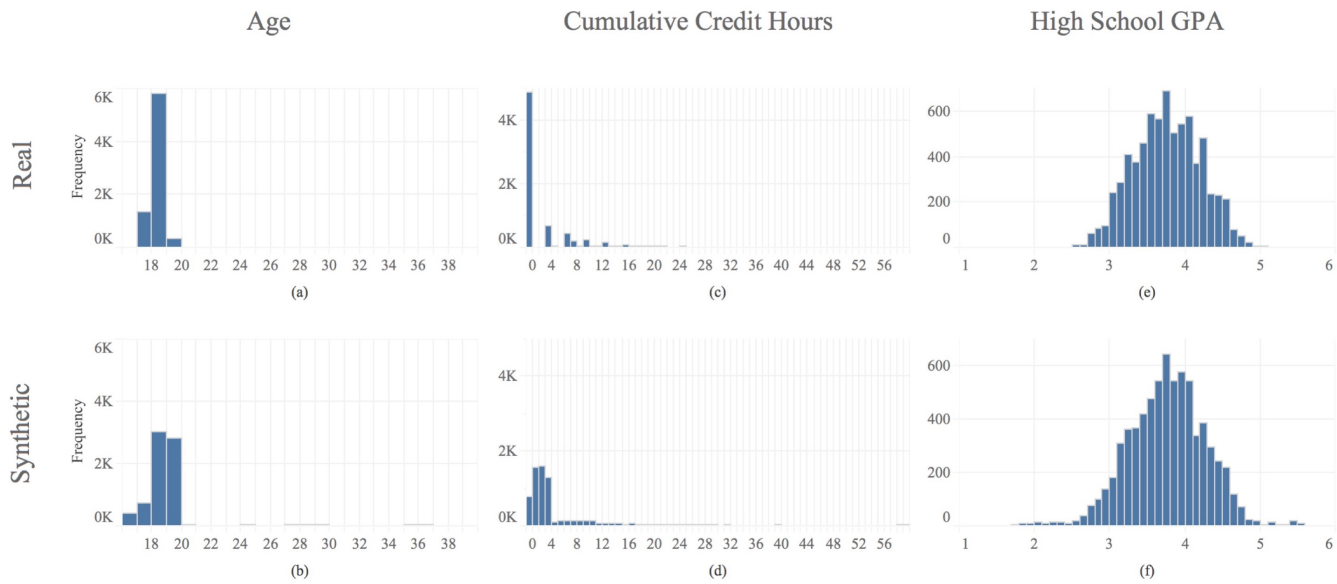

Fig. 1. Comparison of Real and Synthetic distributions for selected features. Bars represent counts of students. (a) Real Age and (b) Synthetic Age. Age (in years) was calculated as of the matriculation date. (c) Real Cumulative Hours Earned and (d) Synthetic Cumulative Hours Earned. These are credit hours earned as of matriculation, including AP credits and any other transferable credits from an institute of higher learning. (e) Real High School GPA and (f) Synthetic High School GPA. High school GPA is on a weighted scale of 6.0.

matriculation date) in Fig. 1 (a and b), cumulative credit hours in Fig. 1 (b and c), and high school grade point average in Fig. 1 (d and e). We notice that in each of these features the distribution of the synthetic data does closely resemble that of the real data. We generated similar figures for each of the hand-selected 20 features in the dataset they all demonstrate a close and similar resemblance. To save space, we only report the features mentioned above that are not institution-specific and generic enough.

Algorithm 1 Data Synthesis

Data: Original Data
Result: Synthetic Data
begin parameter initialization
  $thresholdValue$ = 20
  $epsilon$ = 0.1
  $degreeOfBayesianNetwork$ = 2
  $sampleToGenerate$ = 7206
  $mode$ = $correlatedAttributeMode$
end parameter initialization
Instantiate $DataDescriber(thresholdValue)$
Describe dataset in $mode$
Save dataset description
Instantiate $DataGenerator()$
Generate synthetic dataset in $mode$
Save synthetic dataset

To demonstrate the effect of using the shareable synthetic dataset over real data in machine learning and classification tasks, we designed and ran an experiment. We split both the real and synthetic datasets into 66% train and 33.33% test. In this experiment, we applied the *Scikit-Learn* RandomForest Classifier [20] on both the synthetic and real datasets. We used the following parameters: n_estimators (trees) were 300, max_depth = 30, and random_state = 3. These parameters were kept constant throughout the experiment. The results of the classification task are summarized in table I using accuracy, type 1, and type 2 errors as performance indicators for the classification task. Even though the results on real data seem to be better, the difference does not seem to be quite significant.

TABLE I
CLASSIFICATION COMPARISON

Dataset	*Accuracy%*	*Type I Error%*	*Type II Error%*	*F1-Score*
Real Data	80.8	15.5	3.5	0.79
Synthetic Data	74.2	19.75	6.0	0.72

IV. DISCUSSION AND FUTURE WORK

Learning analytics collects, reports, and studies learners in various ways. Open data and computer code accessibility are crucial for the field to thrive. Student privacy concerns discourage sharing student data. As mentioned, data synthesis has already been used as a solution in various fields to enable data access for researchers while upholding source privacy. In this paper, we proposed using data synthesis and generation as a solution to the data shareability issue in the field of learning analytics. We report a strong resemblance between original and synthesized data and our initial findings indicated potential

reliability for the synthetic data to be used for predictive classification in learning analytics. Through this paper, we call for learning analytics researchers to use data synthesis to increase data accessibility and collaboration.

The use of synthetic datasets has many advantages for higher education research. For instance, enrolled student demographics and characteristics shift over time at many institutions. This complicates applying past lessons learned from historical student data to currently enrolled students. Synthetic datasets allow researchers to evaluate how changes to specific features may influence analytic models, facilitating more concrete understandings about how changes to student characteristics may influence modeling of student data. Synthetic datasets also allow researchers to examine how noise and outliers may impact model performance, which is useful when researchers are interested in applying models to smaller sample sizes.

Resolving the issue of data shareability in learning analytics helps lay the groundwork for important field endeavors such as standard benchmark datasets and increased multi-institution collaboration. Synthetic datasets offer distinct advantages for higher education researchers working on learning analytics, and may help promote adoption of open science's best practices by the learning analytics field.

REFERENCES

[1] M. Woelfle, P. Olliaro, and M. H. Todd, "Open science is a research accelerator," *Nature Chemistry*, vol. 3, no. 10, p. 745, 2011.

[2] R. Vicente-Sáez and C. Martínez-Fuentes, "Open science now: A systematic literature review for an integrated definition," *Journal of business research*, vol. 88, pp. 428–436, 2018.

[3] S. Fiore, D. Elia, C. Palazzo, A. D'Anca, F. Antonio, D. N. Williams, I. Foster, and G. Aloisio, "Towards an open (data) science analytics-hub for reproducible multi-model climate analysis at scale," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 3226–3234.

[4] G. Popkin, "Setting your data free," *Nature*, vol. 569, no. 7756, pp. 445–447, 2019.

[5] M. A. Chatti, A. Muslim, and U. Schroeder, "Toward an open learning analytics ecosystem," in *Big data and learning analytics in higher education*. Springer, 2017, pp. 195–219.

[6] M. Khalil and M. Ebner, "What is learning analytics about? a survey of different methods used in 2013-2015," *arXiv preprint arXiv:1606.02878*, 2016.

[7] G. Siemens, D. Gasevic, C. Haythornthwaite, S. Dawson, S. B. Shum, R. Ferguson, E. Duval, K. Verbert, and R. Baker, "Open learning analytics: an integrated & modularized platform," Ph.D. dissertation, Open University Press Doctoral dissertation, 2011.

[8] J. T. Avella, M. Kebritchi, S. G. Nunn, and T. Kanai, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review." *Online Learning*, vol. 20, no. 2, pp. 13–29, 2016.

[9] F. E. Rights and P. Act, "Usc 1232-34 cfr part 99," 1974.

[10] H. Ping, J. Stoyanovich, and B. Howe, "Datasynthesizer: Privacy-preserving synthetic datasets," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 2017, p. 42.

[11] Y. Pei and O. Zaïane, "A synthetic data generator for clustering and outlier analysis," 2006.

[12] P. Christen and A. Pudjijono, "Accurate synthetic generation of realistic personal information," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2009, pp. 507–514.

[13] A. M. Berg, S. T. Mol, G. Kismihók, and N. Sclater, "The role of a reference synthetic data generator within the field of learning analytics." *Journal of Learning Analytics*, vol. 3, no. 1, pp. 107–128, 2016.

[14] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016, pp. 399–410.

[15] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd, "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, vol. 79, no. 3, pp. 362–384, 2011.

[16] S. K. Kinney, J. P. Reiter, and J. Miranda, "Synlbd 2.0: improving the synthetic longitudinal business database," *Statistical Journal of the IAOS*, vol. 30, no. 2, pp. 129–135, 2014.

[17] L. Liang, J. Zhong, J. Liu, P. Li, C. Zhan, and Z. Meng, "An implementation of synthetic generation of wind data series," in *2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 2013, pp. 1–6.

[18] E. L. Barse, H. Kvarnstrom, and E. Jonsson, "Synthesizing test data for fraud detection systems," in *19th Annual Computer Security Applications Conference, 2003. Proceedings.* IEEE, 2003, pp. 384–394.

[19] A. Botta, A. Dainotti, and A. Pescapé, "A tool for the generation of realistic network workload for emerging networking scenarios," *Computer Networks*, vol. 56, no. 15, pp. 3531–3547, 2012.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.