Konzeption und Umsetzung synthetischer Datengenerierung für Forschung und Entwicklung in Assessment Analytics

Martin Breuer[1], Malte Persike[1] und Ulrik Schroeder[2]

Abstract: Die Datenbeschaffung für Learning Analytics zur Verbesserung der Bildungstechnologien und Lehrinhalte wird durch lange Wartezeiten, Unklarheiten bezüglich Datenverfügbarkeit und Datenschutz sowie Art der Daten erschwert. Synthetische Datensätze können diese Hürden überwinden, indem sie als Ersatz für echte Daten dienen. In der Literatur werden sowohl Ansätze zur synthetischen Datengenerierung auf Basis echter Daten (z. B. mithilfe künstlicher Intelligenz), als auch Ansätze zur Simulation der Interaktionen auf Basis konkreter Annahmen beschrieben. Diese synthetischen Daten sollen die Entwicklung von Assessment Analytics Tools für das E-Prüfungssystem Dynexite unterstützen. Dieser Beitrag stellt erste Entwicklungsschritte zur Bereitstellung von Testdatensätzen für das E-Prüfungssystem vor: Ein erster Prototyp erstellt automatisiert Prüfungen mit variabler Teilnehmerzahl und zufälligen Ergebnissen. Die notwendigen Konzepte werden in einem nächsten Schritt zu einem Daten-Erstellungs-Service abstrahiert, um die einfache Programmierung weiterer Datensätze zu ermöglichen. Zur Demonstration der Funktionalität wird ein öffentlich verfügbarer Learning Analytics Datensatz importiert. Ein erster Testlauf konnte bereits Skalierbarkeitsprobleme der bestehenden Codebasis aufdecken.

Keywords: E-Prüfungssystem, Assessment Analytics, Learning Analytics, synthetic data.

1 Einleitung

Die Entwicklung und Implementierung skalierbarer Learning Analytics Anwendungen stellt Hochschulen vor technische Herausforderungen. Zentrale Learning Analytics Infrastrukturen wie EXCALIBUR LA [JS22] ermöglichen die Sammlung und Untersuchung von Lerner Daten, die in verschiedenen Plattformen entstehen. Bei der Erweiterung von Learning Analytics Anwendungsfällen und der Anbindung weiterer Plattformen ist es entscheidend, die Verarbeitung realistischer und wachsender Datenmengen zu ermöglichen.

Während der Einsatz effizienterer Verfahren, leistungsstärkerer Maschinen (Scale up) oder Lastverteilung auf mehrere Maschinen (Scale out) erfolgen kann, ist bei der Forschung, Entwicklung und Implementierung lediglich eine Skalierung durch Kollaboration (Scale out) möglich. Die Bereitstellung synthetischer Testdaten ermöglicht es Forschern, Verfahren und Algorithmen zu verbessern und Kooperation zu fördern [Be16].

[1] RWTH Aachen, Center für Lehr- und Lernservices, Kackertstr. 15, 52074 Aachen, {breuer@medien, persike@cls}.rwth-aachen.de, https://orcid.org/{0009-0008-0749-5110, 0000-0002-7825-089X}
[2] RWTH Aachen, Lehr- und Forschungsgebiet Informatik 9, Ahornstr. 55, 52074 Aachen, {a.brocker, schroeder}@informatik.rwth-aachen.de, https://orcid.org/0000-0002-5178-8497

doi: 10.18420/wsdelfi2023-32

Dieser Beitrag soll zunächst Möglichkeiten zur Nutzung synthetischer Datengenerierung erläutern und anschließend eine technische Basis für die Einbindung der generierten Daten für zukünftige Forschungs- und Entwicklungsschritte von Assessment Analytics Tools für das E-Prüfungssystem Dynexite[3] schaffen.

2 Generierte Datensätze für Forschung und Entwicklung in Assessment Analytics

Ellis definiert Assessment Analytics als Spezialisierung von Learning Analytics im Kontext von Assessments [El13]. Learning Analytics ist zunächst das Messen, Sammeln, Analysieren und Auswerten von Daten über Lernende und Ihren Kontext, mit dem Ziel, das Lernen und die Lernumgebung zu verstehen und zu optimieren [SL11].

Knaub et al. untersuchen die Unterstützung von Learning Analytics durch Personal zur Bereitstellung von Lerner Daten und zur Begleitung des Learning Analytics Prozesses, wobei Lehrende befragt wurden. Herausforderungen bei der Nutzung dieser Unterstützungsmöglichkeit umfassen sowohl individuelle Probleme beim Erhalt und der Auswertung der Lerner Daten, als auch prozessabhängige Hürden, wie lange Wartezeiten zum Erhalt der Daten, Unwissenheit, welche Daten verfügbar sind und wie sie genutzt werden können, sowie Bedenken zum Schutz der Lerner Daten. Als Chance wird die Bereitstellung synthetischer Daten in einem Web-basierten Datenexplorer genannt, um herauszufinden, welche Daten verfügbar sind und ob die Beschaffung echter Daten lohnenswert ist, oder ob relevante Daten nicht verfügbar sind. [Kn16]

Berg et al. diskutieren, dass ein synthetischer Datengenerator in frühen Entwicklungsphasen bei der Entwicklung und Optimierung von Prozessen helfen kann, bevor reale Daten verfügbar sind. Insbesondere, da reale Daten durch niedrigere Qualität von Alpha und Beta Software in nicht-produktiv Umgebungen durch unbeabsichtigte Offenlegung stärker gefährdet sind. Synthetische Daten ermöglichen Dienste aufzubauen, bevor politische, ethische, rechtliche und datenschutzrechtliche Probleme geklärt sind. [Be16]

Zur Erzeugung synthetischer Datensätze werden im Folgenden zwei Ansätze betrachtet: Die erste Methode nutzt echte Daten als Grundlage während die zweite mithilfe von Simulationsmodellen unter Verwendung konkreter Annahmen erfolgt.

2.1 Synthetische Datengenerierung

Eine Methode zur Erstellung synthetischer Datensätze, die ähnliche statistische Eigenschaften wie echte Daten besitzen, ohne jedoch vertrauliche Informationen zu replizieren, ist die Verwendung von Generative Adversarial Networks (GANs) [BI21]. Hierbei werden zwei neuronale Netzwerke eingesetzt: eines zur Datengenerierung und ein weiteres

[3] Dynexite Dokumentation, https://docs.dynexite.rwth-aachen.de/, Stand: 27.03.2023

zur Diskriminierung zwischen echten und synthetischen Daten. Beide Netzwerke trainieren und verbessern sich gegenseitig in Konkurrenz. Dieser Ansatz kann einerseits die Privatsphäre der Lernenden (auch in kleineren Kursgrößen) schützen und andererseits die Datensatzgröße durch Erzeugung weiterer Datenpunkte beliebig erhöhen [BI21]. Ein Nachteil ist, dass *Overfitting* zu einer 1:1 Replikation des echten Datensatzes führen kann.

Ein weiterer Anwendungsbereich für die Generierung synthetischer Daten besteht darin, fehlende Werte zu ergänzen und das sogenannte *Minority Oversampling* durchzuführen [Ko18]. Letzteres erstellt synthetische Datenpunkte für unterrepräsentierte Minderheitsklassen, zur Verbesserung der Leistung maschinellen Lernens.

2.2 Simulierte Datengenerierung

Ein weiterer Ansatz zur Generierung synthetischer Daten ist die Simulation von Lerner-interaktionen, ohne auf echte Daten angewiesen zu sein. Hierbei werden Modelle und Annahmen entwickelt, die bestimmten Szenarien oder theoretischen Konzepten entsprechen. In einem systematischen Literaturreview [KA23] von Alexandron und Käser werden verschiedene Anwendungsfälle für die Simulation von Lernerinteraktionen im Bildungsbereich und verwandten Gebieten vorgestellt. In der Praxis konzentrieren sich Simulationsmodelle meist nur auf bestimmte Aspekte des Lernens und es erfolgt meist keine Sicherung der Validität. Die Autoren schlagen ein Turing-ähnliches Validitätskriterium vor, bei dem ein Simulationsmodell valide ist, wenn eine Gruppe von Bildungsexperten keinen Verhaltensunterschied zu echten Lernern feststellen können. Zukünftige Forschung könnte sich auf die Entwicklung vollständigerer Lerner Modelle konzentrieren. [KA23]

Der folgende Abschnitt soll erste Erfahrungen bei der technischen Umsetzung zur Generierung eines Referenzdatensatzes für das Prüfungssystem Dynexite³ vorstellen.

3 Datenimport in Bildungstechnologie

Das Ziel ist die Bereitstellung einer Instanz des Prüfungssystems Dynexite³ inklusive eines Testdatensatzes. Dieser Abschnitt beschäftigt sich mit der Frage, wie ein Datensatz in das Prüfungssystem Dynexite³ importiert werden kann. Ob die Daten synthetisch generiert, simuliert oder aus einem bestehenden Datensatz anonymisiert sind, ist für die Schaffung einer technischen Basis zunächst unbedeutend. Ein konkreter Anwendungsfall ist die Weiterentwicklung und Prüfung der Skalierbarkeit und Validität einer Logging-Erweiterung des Prüfungssystems als Datenquelle für eine Assessment Analytics Anwendung [Br23]. Diese würde Kennwerte wie die Ratewahrscheinlichkeit, Bearbeitungsdauer oder Schwierigkeitsgrad ermitteln [Br23].

Zum Test dieser Funktionalitäten wäre die Bereitstellung eines Datensatzes ideal, der ein kleinschrittiges Verfolgen des Lösungsverhaltens der Studierenden mit Bezug zu einzelnen Eingabefeldern ermöglicht. Als Startpunkt dient der frei verfügbare Open University

Learning Analytics Datensatz (OULAD) [KHZ17], der anonymisierte Ergebnisdaten zu Übungen und Prüfungen sowie die Nutzung von Lernmaterialien von ca. 23.000 Studierende in 22 Kursen mit insgesamt 174.000 Prüfungs- und Übungsversuchen enthält. Der Datensatz wurde aufgrund schneller Verfügbarkeit, verständlicher Dokumentation, simplem Dateiformat (CSV) sowie auf den ersten Blick hoher Popularität ausgewählt. Er enthält zwar keinen feingranularen Änderungsverlauf der Antworten während der Prüfungen, diese könnten zu einem späteren Zeitpunkt jedoch synthetisch ergänzt werden, um Testfälle für Detailauswertungen zu entwickeln.

Damit die Prüfungen eines solchen Datensatzes wie echte Prüfungen in Dynexite[3] sichtbar sind und als Datenquelle für Assessment Analytics dienen können, müssen sowohl die Dynexite-Event-Logs, als auch der aktuelle Zustand der relationalen Datenbank des Prüfungssystems repliziert werden. Eine vollständige Simulation der Benutzerinteraktionen in der Web-Oberfläche des Prüfungssystems wäre eine realistische Methode zur Generierung des Testdatensatzes. Erfahrungen aus früheren Projekten zeigen jedoch, dass der anfängliche Aufwand für die Erstellung der Simulation und der Wartungsaufwand nach Aktualisierungen der Benutzeroberfläche hoch sind. Die Nutzung der Backend-REST-Schnittstelle würde den Aufwand für Entwicklung und Wartung bereits reduzieren. Die hexagonale Softwarearchitektur des Prüfungssystems, trennt jedoch bereits Businesslogik nach außen verfügbaren Schnittstellen [Co05]. Diese Trennung erlaubt die direkte Verwendung der Businesslogik, sodass Services zur Erstellung von Aufgaben, zum Starten einer Prüfung usw. direkt aufgerufen werden können.

3.1 Durchführung einer Prüfung mit 100 Studierenden

Als erster Schritt vor der Bereitstellung realistischer Prüfungsdaten, gilt es zunächst eine Prüfung maschinell zu erstellen und die Durchführung mit mehreren simulierten Studierenden (z. B. 100) zu imitieren. Aus technischer Sicht muss eine Prüfung mindestens eine Aufgabe enthalten und zum Beenden der Korrekturphase muss jeder Teilnehmende eine Bewertung (Punktzahl) für diese Aufgabe erhalten. Hierzu wird ein einfacher Zufallszahlengenerator verwendet. Namen für Aufgaben, Kurse etc. werden zufällig erstellt, um Duplikate bei wiederholter Ausführung zu vermeiden.

Eine erste lauffähige Version kann mit wenig Entwicklungsaufwand erstellt werden, indem zentrale Komponenten des Prüfungssystems wiederverwendet werden. Da lediglich die Nutzerinteraktion ersetzt wird, können zentrale Konfigurationen wie die Datenbankverbindung beibehalten werden. Die Services zur Erstellung von Aufgaben, Prüfungen, Studierenden usw. können anschließend in einen Service zur Datengenerierung eingebunden werden. Der Prüfungsprozess wird schrittweise durch Aufrufe der Servicefunktionen, z. B. *ItemCreator.CreateItem(…)*, simuliert.

Während dieses Entwicklungsschrittes hat sich gezeigt, dass die bestehenden Komponenten des Prüfungssystems mit wenig Aufwand neu zusammengesetzt werden können, um Codegetrieben Daten zu generieren, die vom Prüfungssystem genauso behandelt werden

wie echte Daten. Als Nebenprodukt konnten mögliche Arbeitspakete zur Verbesserung der Codequalität des Prüfungssystems ermittelt werden, die in der Planung zukünftiger Versionen des Prüfungssystems berücksichtigt werden: in seltenen Fällen gab es Businesslogik außerhalb der Services, teilweise unregelmäßige Benennungen von Dateinamen oder dass Zustandsänderungen zwar gespeichert wurden, zur Weiterverarbeitung aber erneut aus der Datenbank geladen werden mussten.

3.2 Import des Open University Learning Analytics Datensatzes [KHZ17]

Im vorherigen Abschnitt wurde die Simulation des Prüfungsprozesses ohne realistische Daten vorgestellt. Nun erfolgt die Importierung des Open University Learning Analytics Datensatzes (OULAD) [KHZ17]. Um den Import dieses Datensatzes und weiterer synthetischer Datensätze zu erleichtern, wurde die Lösung des letzten Abschnitts refaktoriert. Wiederverwendbare Methoden zum Erstellen von Studierenden, Kursen, Kurszugehörigkeiten, Prüfungsversuchen, Bewertungen usw. wurden in einem Service als weitere Abstraktionsschicht gebündelt. Nach dem Auslesen der CSV-Dateien des OULA-Datensatzes konnten diese Methoden zur erfolgreichen Importierung der Prüfungs- und Übungsversuche inklusive Kurs- und Pseudonym Zuordnung dienen. Eine Limitierung des Datenimports besteht aktuell darin, dass nur Ergebnisdaten importiert wurden. Um zukünftig feingranulare Event-Daten mit realistischen Zeitstempeln zu simulieren, ist die Bereitstellung einer manipulierbaren Uhr im Prüfungssystem notwendig, damit Interaktionen nicht die aktuelle Uhrzeit des Datenimports, sondern des simulierten Zeitpunktes erhält. Dies ist voraussichtlich einfach umsetzbar, jedoch nicht Teil dieses Datenimports, da die bestehende Codebasis des Prüfungssystems angepasst werden müsste.

Ein erster Testdurchlauf unterstreicht die Vorteile eines Testdatensatzes mit realistischer Größe (z. B. > 1000 Teilnehmende): Der Import der ersten Prüfungen dauerte ungewöhnlich lange. Die Untersuchung der Event-Logs deutete auf ein Skalierungsproblem im Prozessschritt zur Beendigung der Ausführungszeit der Prüfungen hin. Durch die frühzeitige Erkennung konnte das Problem vor dem Produktiveinsatz der Prüfungssystem-Version behoben und der Datensatz ohne auffällige Skalierbarkeitsprobleme importiert werden.

4 Fazit

Zusammenfassend lässt sich sagen, dass sowohl die synthetische Datengenerierung auf Basis echter Daten, als auch die Entwicklung von Simulationsmodellen ein breites Anwendungsspektrum bieten. Obwohl beide Möglichkeiten jeweils keine One-Size-Fits-All Lösung für Datenschutz (z. B. durch Overfittung) oder die Erstellung realistischer Daten bieten können, scheint der Einsatz in Forschung und Entwicklung im Bildungsbereich vielversprechend.

Der vorgestellte Datenimport ermöglicht das detaillierte Betrachten der Daten in Korrektur- und Ergebnisansicht des Prüfungssystems und das Testen grobgranularer Assessment

Analytics Untersuchungen (vgl. [BR23]). Im Rahmen des Projektes NOVA:ea, gefördert durch die Stiftung Innovation in der Hochschullehre, wird ein Prüfungscockpit zur iterativen Verbesserung der E-Prüfungen entwickelt. Die hier vorgestellte technische Basis dient zur Bereitstellung synthetischer Daten für Forschung und Entwicklung. Synthetische Daten sollen künftig auch verwendet werden, um Datenpunkte zu ergänzen, die in aktuell verfügbaren Daten nicht enthalten sind.

Literaturverzeichnis

[BI21] Bautista, P.; Inventado, P. S.: Protecting Student Privacy with Synthetic Data from Generative Adversarial Networks. In (Roll et al.): Artificial Intelligence in Education, Lecture Notes in Computer Science. Bd. 12749. Springer International Publishing, Cham , S. 66–70, 2021.

[Br23] Breuer, M. et al.: AxEL - Eine modulare Softwarekomponente für ein dediziertes E-Prüfungssystem zur Generierung von xAPI Statements für Assessment Analytics : 20. Fachtagung Bildungstechnologien (DELFI), S. 91-96 , 2023.

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

[Co05] Cockburn, A.: Hexagonal architecture, 2005. URL https://alistair.cockburn.us/hexagonal-architecture/. - abgerufen am 2023-10-02.

[El13] Ellis, C.: Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. In: British Journal of Educational Technology, S. 662-664, 2013.

[JS22] Judel, S.; Schroeder, U.: EXCALIBUR LA - An Extendable and Scalable Infrastructure Build for Learning Analytics. In: 2022 International Conference on Advanced Learning Technologies (ICALT), S. 155–157, 2022.

[KA23] Käser, T.; Alexandron, G.: Simulated Learners in Educational Technology: A Systematic Literature Review and a Turing-like Test. In: International Journal of Artificial Intelligence in Education, 2023.

[Ko18] Kovanović, V. et al.: Understand students' self-reflections through learning analytics. In: Proceedings of the 8th International Conference on Learning Analytics and Knowledge. ACM, Sydney New South Wales Australia, S. 389–398, 2018.

[Kn16] Knaub, A. V. et al.: Supporting faculty and staff to make better use of learning analytics data. In: 2016 Physics Education Research Conference Proceedings, American Association of Physics Teachers, Sacramento, CA, S. 188–191, 2016.

[KHZ17] Kuzilek, J; Hlosta, M.; Zdrahal, Z.: Open University Learning Analytics dataset. In:

Scientific Data Bd. 4, Nature Publishing Group, Nr. 1, S. 170-171, 2017.

[SL11] Siemens, G.; Long, P.: Penetrating the fog: Analytics in learning and education. In: EDUCAUSE review 46(5), S. 31-40, 2011.