Einsatz von GANs und LLMs zur Erzeugung von synthetischen Daten in der Pharmakogenetik

Studiengang: BSc in Medizininformatik
Betreuer: Prof. Dr. Murat Sariyar

Synthetische Daten sind künstlich erzeugte Daten, die ähnliche statistische Eigenschaften wie reale Daten aufweisen und daher für eine Vielzahl von Anwendungen verwendet werden können. Ziel dieser Arbeit war es, mithilfe von GANs (Generative Adversarial Networks) und einem LLM (Large Language Model) synthetische Daten für die Beobachtungsstudie der Pharmaceutical Care Research Group der Universität Basel zu generieren und deren Qualität zu bewerten.

Einführung

Die Pharmakogenetik (PGx) untersucht den Einfluss der individuellen genetischen Disposition auf die Wirksamkeit und Verträglichkeit von Arzneimitteln. Die Pharmaceutical Care Research Group der Universität Basel hat zu diesem Thema eine Beobachtungsstudie gestartet. Im Rahmen der Studie werden Daten von Patienten gesammelt, die unerwünschte Arzneimittelwirkungen oder Therapieversagen erleiden. Um den Bedarf an zusätzlichen PGx-Daten für die oben genannte Beobachtungsstudie zu decken, wurden im Rahmen des Projektes synthetische Daten mithilfe von drei Methoden generiert. Dazu wurden zwei GANs (Generative Adversarial Networks), CT-GAN und CTAB-GAN+, sowie ein LLM (Large Language Model) namens Tabula verwendet.

Methode

GANs sind ein maschinelles Lernmodell, das aus zwei neuronalen Netzen besteht: einem Generator, der Daten generiert, und einem Diskriminator, der diese von realen Daten unterscheiden soll. Über mehrere Iterationen lernt der Generator, die statistischen Eigenschaften der Originaldaten zu imitieren, während der Diskriminator immer besser synthetische von realen Daten unterscheiden kann. Im Rahmen dieses Projekts werden zwei Arten von GANs verwendet, die den beschriebenen Mechanismus leicht modifizieren: CTGAN und CTAB-GAN+. Zum Vergleich wurde ein sogenanntes Large Language Model (LLM) verwendet. Dieses versucht, Zusammenhänge in der Sprache zu erkennen und kontextangepassten Text zu generieren. Obwohl LLMs in erster Linie für die Textgenerierung entwickelt wurden, können sie auch für die Erzeugung von tabellarischen Daten verwendet werden. Mit Hilfe von Korrelationsmatrizen wurde überprüft, ob Zusammenhänge zwischen den Spalten der Datensätze erhalten geblieben sind. Mit einem weiteren Testverfahren (Identifiability) wurde sichergestellt, dass die synthetischen Daten den Originaldaten nicht zu ähnlich sind und keine Rückschlüsse auf die Originaldaten zulassen. Zusätzlich wurde untersucht, wie gut sich Klassifikationsaufgaben, wie die Vorhersage einer Medikationsänderung, mit synthetischen Daten lösen lassen.

Resultate

Im Gegensatz zu den GANs hatte das LLM Probleme mit der Datengenerierung. Die Anzahl der Spalten musste für das LLM stark reduziert werden. Generell wurde durch die angewandten Verfahren eine hohe Ähnlichkeit mit den Originaldaten erreicht. Es zeigte sich, dass die synthetischen Daten einen hohen Anonymisierungsgrad aufweisen, wobei CTGAN hier die besten Werte erzielte. Zudem wurde festgestellt, dass die synthetischen Daten unabhängig von der Generierungsmethode eine ähnliche Leistung wie die Originaldaten erbringen. Bei der Klassifikation realer Daten schnitten Modelle, die mit einer grossen Anzahl synthetischer Daten trainiert wurden, leicht besser ab als Modelle, die mit realen Daten trainiert wurden.

Diskussion

In diesem Projekt hat sich gezeigt, dass PGx-Daten besondere Eigenschaften aufweisen, die eine synthetische Datengenerierung erschweren. Dies liegt zum einen an der geringen Anzahl von Beobachtungen und zum anderen an der hohen Anzahl von Spalten. Aus diesem Grund war das LLM nicht in der Lage, komplexe Datensätze zu nachzubilden. Obwohl das LLM sehr gute Daten mit Datensätzen erzeugen kann, die über ein geringe Anzahl Spalten verfügen, ist das Verfahren im Gegensatz zu GANs für PGx-Daten weniger geeignet. Bei den GANs wies CTABGAN+ insgesamt eine höhere Ähnlichkeit mit den Originaldaten auf als CTGAN und eignet sich somit besonders gut für den vorliegenden Use Case.

Dominik Andrej Aeschbacher
Advanced Data Processing

Jessica Meisner
Advanced Data Processing