Anonymous Big Data

AI-Based Privacy-Preserving Big Data Sharing for Market Research

Programm / Ausschreibung	IKT der Zukunft, IKT der Zukunft, IKT der Zukunft - 6. Ausschreibung (2017)	**Status**	abgeschlossen
Projektstart	01.10.2019	**Projektende**	28.02.2022
Zeitraum	2019 - 2022	**Projektlaufzeit**	29 Monate
Keywords	anonymization; data protection; market research; artificial intelligence; big data		

Projektbeschreibung

Offener Datenaustausch hat das Potential den wissenschaftlichen Fortschritt zu beschleunigen und positive Wohlfahrtseffekte zu generieren. Gleichzeitig ist der Schutz personenbezogener Daten ein Grundrecht in der Europäischen Union, welches kürzlich durch die Datenschutz-Grundverordnung (DSGVO) vereinheitlicht und gestärkt wurde. Aus diesem Grund sehen sich Organisationen mit strengeren Regeln für die Verarbeitung und den Austausch von personenbezogenen Daten konfrontiert (siehe Appl et al., 2017).

Anonymisierungsverfahren stellen eine Lösung dieses Zielkonfliktes dar, indem sie einerseits sicherstellen, dass Daten nicht auf Einzelpersonen zurückverfolgt werden können, andererseits den Informationsgehalt der Daten weitestgehend erhalten. Herkömmliche Anonymisierungsverfahren, wie Generalisierung, Relabeling, Verzerrung und / oder Aggregation, scheitern jedoch bei Vorliegen von hochdimensionalen, stark korrelierten Daten (siehe Narayanan und Shmatikov, 2008; De Montjoye et al., 2013). Generative mehrschichtige neuronale Netzwerke bieten sich hier als Alternative an. Als brandaktuelles Forschungsgebiet innerhalb der künstlichen Intelligenz weisen diese eindrucksvolle Ergebnisse im Bereich der synthetische Bilderzeugung auf (Karras et al., 2017). Gleichzeitig werden Methoden entwickelt, um sicherzustellen, dass die Information der einzelnen Trainingsdaten im finalen Deep Learning Modell limitiert werden kann (Abadi et al., 2016). Einige wenige Arbeiten verbinden diese beiden Forschungslinien bereits und zeigen auf wie neuronale Netzwerke verwendet werden können, um synthetische Daten zu generieren, welche dann als Ersatz für die originären, personenbezogenen Daten für weitere statistische Analysen verwertet werden können (Beaulieu-Jones et al., 2017).

Ziel dieses Forschungsprojektes ist es generative neuronale Netzwerkarchitekturen für sequenzielle, personenbezogene Daten zu trainieren, um anschließend systematisch zu validieren, inwiefern die Verwendung solcher synthetischer, datenschutzkonformer Daten für die Marketingforschung Dritter nutzbar sind. Von einer derartigen Lösung sollte auch der Datenmarkt Österreich profitieren, da sich schlagartig neue Daten-Services eröffnen würden. Zu diesem Zweck werden zunächst Anforderungen und Anwendungsfälle von personenbezogenen, sequentiellen Daten gesammelt und analysiert. Basierend auf diesen Erkenntnissen wird eine umfangreiche Simulationsstudie mit Hilfe eines eigens entwickelten Virtual Data Labs umgesetzt und anschließend der Ansatz end-to-end mit tatsächlichen empirischen Anwendungsfällen

systematisch validiert.

Angesichts der jüngsten Forschungserfolge von generativen neuronalen Netzwerken für die Synthetisierung von Bildern sind ähnliche Ergebnisse bei deren Anwendung auf sequentielle personenbezogene Daten zu erhoffen. Dies würde maßgeblich zur Lösung einer der größten Herausforderungen Europas im digitalen Zeitalter beitragen, welche in der ökonomischen Nutzung von in personalisierten Längsschnittdaten enthaltenen Informationen bei gleichzeitiger Wahrung der Privatsphäre von Individuen besteht.

Abstract

Open data sharing enables faster scientific progress and creates positive welfare effects for the society. At the same time, the protection of personal data is a fundamental right in the European Union, which recently has been strengthened and unified via the General Data Protection Regulation (GDPR). As a result, organizations face stricter rules on processing and sharing privacy-sensitive data (see Appl et al., 2017).

Ideally, anonymization of privacy-sensitive data offers a solution by ensuring that data cannot, by any means, be traced back to individuals, while retaining most of its utility. However, current anonymization techniques, such as generalization, obfuscation, perturbation, and aggregation fall short in the presence of high dimensional, highly correlated data, which arise when observing individuals over a period of time, i.e. for sequential personal data (see Narayanan and Shmatikov, 2008; De Montjoye et al., 2013).

Generative deep neural networks have recently become a highly active research field within artificial intelligence, with impressive demonstrations for synthetic image generation (see Karras et al., 2017). Further, methods are being developed that allow capping the individual-level information, and thus provide formal privacy guarantees for training neural networks (see Abadi et al., 2016). A first few papers put these two streams already together, showcasing how privacy-preserving deep neural networks can be used to generate synthetic data, that is shared in lieu of actual privacy-sensitive data (see Beaulieu-Jones et al., 2017).

The goal of this research is to, for the first time, train deep generative model architectures to sequential personal data while providing differential privacy guarantees, in order to systematically validate the feasibility of using synthetic, privacy-preserving sequential data for third party market research. Since such a solution would offer tremendous opportunities for completely new data services such a solution should also be extremely beneficial for the Austrian Data Market.

For that purpose, requirements and use cases of privacy-sensitive sequential data will be collected and analyzed. Based on these findings, a virtual data lab is designed and created, that allows to systematically investigate the conditions under which a variety of deep generative models are able to derive synthetic replicas which capture structure and correlations, while protecting individual-level privacy. Ultimately, the approach is then validated end-to-end with actual empirical use cases.

Given the very recent leaps in the field of image synthesization, similar advances are to be expected when combining and transferring these developed methods to sequential personal data. This would provide a viable solution to one of Europe’s biggest economic challenges in the digital age, i.e. the utilization of the growing asset of personal data, while safeguarding

the privacy of individuals.

Projektkoordinator

- Wirtschaftsuniversität Wien

Projektpartner

- George Labs GmbH
- Mostly AI Solutions MP GmbH
- Bundesanstalt "Statistik Österreich"