Medical Xpress

Measuring differential privacy could balance meaningful analytics and health care data security

May 6 2024

Tombs and Bridges have developed a new method that improves on the standard method of differential privacy to allow health care data sharing while maintaining patient privacy. Credit: ORNL, U.S. Dept. of Energy

In industries such as health care, where data generation grows by 47% each year, information collected within electronic health records could help inform more efficient care operations or more accurate diagnoses. However, personal health data is highly protected and largely goes untouched by analysts and researchers.

To balance personal safety and research innovation, researchers at the Department of Energy's Oak Ridge National Laboratory are employing a mathematical technique known as differential privacy to provide data privacy guarantees.

Differential privacy is a measure of how private a mechanism—such as an equation, an algorithm or an analysis—is when applied to a dataset to produce an output. This can help measure how much any one piece of data could affect the output of a dataset, and therefore, be identified. To strengthen differential privacy, researchers add noise, or randomness, to datasets.

ORNL is a perfect testbed for innovating privacy research, providing both high-performance computing resources and a wide range of data that could benefit from differential privacy applications. Researchers in ORNL's National Security Sciences Directorate have partnered with peers in the Computing and Computational Sciences Directorate to advance science while maintaining data privacy.

What puts the 'different' in differential privacy?

Data collected from individuals could reveal trends otherwise unknown to researchers, such as traffic patterns in rideshare apps and disease within certain populations. Removing personal identifying information from aggregate data might not be enough to maintain anonymity. According to Vandy Tombs, an applied mathematician in ORNL's Spatial Statistics group, removing names and personal identifiers to anonymize data isn't enough to de-identify individuals in a dataset.

Tombs is working with Robert (Bobby) Bridges, a research scientist in ORNL's Cyber Resilience and Intelligence Division, to develop a new method of differentially private machine learning.

It turns out, Tombs said, that removing names and personal identifiers to anonymize data isn't enough to de-identify individuals in a dataset.

"Identifying information can be much more than just your name or date of birth," Tombs said. "When there are other pieces of information available, seemingly innocent data—like the day you gave birth to a child—can suddenly become identifying information."

Privacy at the cost of accuracy

The privacy afforded through differential privacy methods comes at a cost of output accuracy. In one example from Tombs, a model trained without privacy on a dataset had an accuracy output of 70%. Applying the same model to the same data set but using the state-of-the-art method for differential privacy training achieved only 20% accuracy with a desired level of privacy. While groundbreaking in the privacy space, the method has its faults, losing privacy at each step.

Tombs and Bridges are testing a new method of differentially private machine learning that has shown promise in maintaining accuracy while increasing privacy over the DPSGD model.

"We are trying to dethrone the king because the king isn't a good king," Bridges said. "(DPSGD) is a very wasteful model."

To usher in a new era, Bridges and Tombs combined techniques from both the differential privacy and machine learning communities. Their new method builds an output distribution over the desired machine learning model in a clever way. The high likelihood models are very accurate, and the distribution is designed to provide strong differential privacy.

This method, called the Exponential Mechanism, was discovered in 2007 but has been sidelined from use in training machine learning models because of one big problem—it is currently impossible to sample from this output distribution. Meanwhile, the machine learning community has developed new techniques for creating tractable approximate distributions.

The hypothesis from Bridges and Tombs is the combination of these methods and may prove transformative for private machine learning. Thus far, their work has shown that sampling a model from the approximate distribution provides much better accuracy for similar privacy. The next step is mathematically proving the privacy guarantee of the approximate output distribution is close to the true privacy.

Adoption at the speed of trust

As with all new technologies and models, the researchers' new approach to differential privacy needs to prove itself on real data to gain traction. The duo began this quest close to home, partnering with Heidi Hanson, a biostatistician in the Advanced Computing in Health Sciences Section of ORNL's Computing and Computational Sciences Directorate. Hanson partners with health care institutions and subject matter experts to bring computational science solutions into the clinic.

One project under Hanson's purview is helping health care centers share childhood cancer data as part of the National Childhood Cancer Registry. While data sharing could prove useful across health care in general, it is of particular significance in the childhood cancer space, Hanson said.

"Childhood cancers are rare," she said. "A single institution has a very hard time doing robust research using the limited amounts of data from their patients, as large datasets are required for statistical power and machine learning methods."

On the flip side, Hanson said, institutions are also hesitant to openly share data in a centralized system. While there are decentralized machine learning methods to address this, such as federated learning, the low levels of data available for rare cancer types hamstring these efforts.

"If you're adding noise to only a few cases, you take quite a hit in accuracy to get to the level of privacy you might need in some of these spaces," Hanson said.

The new differentially private model training seeks to allow institutions to maintain privacy with much less sacrifice of accuracy. The team delivered promising experimental results on National Cancer Institute cancer survival data using ORNL's CITADEL framework to comply with federal health care privacy laws.

Hanson presented these results at the International Childhood Cancer Data Partnership meetings, part of the G7 Cancer Initiative in Paris, France, in November 2023.

Widely deploying this new method of differentially private machine learning could help enable real-time cancer surveillance, matching patients to appropriate clinical trials, and help predict cancer outcomes.

"Within the next three years, we would really like (this method) to be widely used by individuals in the medical community so they can solve big problems, especially for rare diseases like childhood cancer," Hanson said.

Provided by Oak Ridge National Laboratory

Citation: Measuring differential privacy could balance meaningful analytics and health care data security (2024, May 6) retrieved 26 June 2024 from https://medicalxpress.com/news/2024-05-differential-privacy-meaningful-analytics-health.html