_Privacy-Aware Data Integration

Antheas Kapenekakis @ Aalborg University, antheas@cs.aau.dk

Overview

Data privacy as a topic has entered the mainstream conscious. Years of negligent data storage, leading to breaches, and **inappropriate sharing** have made the public reluctant to share their data. In addition, this caused legislators to enact **restrictions** in the **use and handling of data**.

Data, through analysis, has a large **utilitarian and economic value**. It is, therefore, imperative, that workflows for **data sanitation** and **responsible sharing** are created, compliant with current regulations. Those will, in turn, help curb **reluctance** of sharing data.

Data Release Types

Data releases can be broadly split into 2: **datasets** and and the **results of analyses** of the data.

Publishing the **results of an analysis** (models, statistics, census data) lowers the possible exposure of the participants but **limits the usability** of the data to the **questions answered by the release**. However, since the information output is limited it is **easier to protect** the data using a tool such as **Differential Privacy**.

A more general **data release**, such as a **dataset**, allows for answering any arbitrary question. This larger scope raises questions about **which information** should be kept in the data, and how it should be **evaluated in terms of privacy and accuracy**.

Ensuring Privacy and Accuracy

In order to ensure that a **data release** meets **accuracy** and **privacy requirements**, two approaches are used.

First, when modeling the data, great care is taken to create **a compressed representation** (a dense **bayesian network** or a small **neural network** with transfer learning). This limits the information extracted from the data, lowering the **noise** added. Then, **Differential Privacy** is applied to the model, which mathematically bounds the privacy and accuracy output.

The end result is **benchmarked** by measuring the **deviation of queries**, the behavior of **models**, and a suite of **marginal metrics** (statistics based).

Future goals

Test viability of **Differential Privacy** in **Data Synthesis**, create multiple models for **synthesizing multimodal data**, and create a suite of **viable metrics** for that data.

Privacy Aware Data Synthesis

Grab Original Data

Patients

ID	Name	Surname	Gender	Region	Age	Age of Death
1	John	Doe	M	us-west	36	56
2	Mary	Taylor	F	us-east	25	-
3	Andy	Smith	M	us-west	55	-
4	Mark	Anderson	M	us-west	72	77

De-identify
- Remove PHI (ex. names)
- Remove unusable columns
- Tweak Dates: 2/1/23->9/2/78

Admissions

ID	PID	Admit Time	Disch. Time
5	1	3/1/22 12:12	4/1/22 14:14
7	1	7/5/20 14:32	8/5/20 15:32
8	2	3/3/19 09:00	3/3/19 12:10
5	2	...	...
9	2	...	...

Diagnoses

ID	AID	Priority	ICD9
1	5	1	64891
2	5	2	V0251
3	5	3	2825
4	2	1	V270
5	3	1	66411
6	4	1	V0251

Transform
- Make dates to (age, week, day)
- Normalize numeric columns
- Encode categorical variables

Fit Model → Reverse

Patient (1D vector)

Gender bool	Age norm	Death is null	Death norm
0	0.3	0	0.56

Diagnoses per admission (2D Matrix)

Disease Priority	has_disease 1D bool	icd9_chapter one hot (1x17)	icd9_cat * base int	icd9_subcat * base int
0	1	001000...	3	2
1	1	000010...	2	3
2	1	100000...	5	0
...	0	000000...	0	0
6	0	000000...	0	0

Patient Admissions (1D vector per)

Admit Time	Disch. Time
(38, 1, 0) 12:12	(38, 1, 2) 14:14
(36, 19, 3) 14:32	(36, 19, 4) 15:32
normalizes to (%f, %f, one-hot, %f):	
0.38, 0.02 0000001, 0.51	0.38, 0.02 0000100, 0.6
0.36, 0.4, 0001000, 0.7	0.36, 0.4, 0010000, 0.74

Testing of Synthetic Data

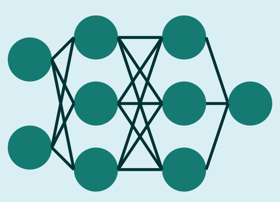
CS	X^2	p
lang	0.01	0.89
ethn	0.03	0.7

KS	KS	p
age	0.02	0.78
death	0.05	0.68

KL	diag	meds
age	0.8	0.78
death	0.7	0.5

AVG err: **select avg**(age) **from** patients
select count(*) **from** diagnoses **group by** icd9 **limit** 20

Train classifiers, **Test** accuracy:



↑ $\text{accuracy}_{\text{real data}}$ = ↑ quality
$\text{accuracy}_{\text{train data}} > \text{accuracy}_{\text{test data}}$ = ↓ privacy

Aalborg University
Katja Hose, Daniele Dell'Aglio

Athena Research Center
Minos Garofalakis

DEDS