VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Bakalauro baigiamojo darbo planas

Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas

(Privacy-preserving artificial intelligence)

Atliko: 4 kurso 3 grupės studentas
Paulius Milmantas (parašas)
Darbo vadovas:
dr. Linas Petkevičius (parašas)

Vilnius
2021

Turinys

Įvadas

Mašininis mokymas yra dirbtinio intelekto sritis, kuri pasitelkia statistinius algoritmus, kad apibrėžtų duomenų generavimo mechanizmą, ar egzistuojančius sąryšius, priklausomybes. Modelis dažnai turi didelį kiekį nežinomų parametrų, kuriuos reikia įvertinti pagal duomenis, todėl modelio apmokymui dažniausiai reikia turėti daug duomenų. Kai kurie uždaviniai reikalauja duomenų, kurie nėra laisvai prieinami ir yra privatūs. Mašininio mokymo tyrimų srityje yra kilusi problema dėl jų saugojimo [BJJ+18]. Vienas iš faktorių, kuris lėmė šį susidomėjimą yra 2016 metais Europos Sąjungoje priimtas duomenų apsaugos reglamentas (GDPR). Pagal jį, fizinių asmenų duomenys turi būti saugomi naudojantis tam tikromis taisyklėmis ir negali būti atskleisti trečiosioms šalims be asmens sutikimo [116].

Šią problemą išspręsti siekia įvairūs tyrimai ir naujai pasiūlyti metodai privatumą saugančio dirbtinio intelekto srityje. Šią problemą galima išskaidyti į kelias atskiras sritis:

- Analizuojamų duomenų privatumas [Tha20]. Algoritmas apmoko modelį atpažinti duomenis. Turint sukurtą modelį, neturi būti galima atgaminti duomenų, pagal kuriuos jis buvo mokomas, bei negali būti identifikuoti asmenys. Taip nukentėtų žmonių privatumas ir būtų pažeistas Europos duomenų apsaugos reglamentas. Šio pažeidimo pavyzdys gali būti ir paprastas teksto atkūrimo modelis. Duodama sakinio pradžia, modelis nuspėja jo pabaigą. Jeigu suvedus tam tikras detales modelis užbaigia sakinį naudodamas asmeninius duomenis, kurie atskleidžia žmonių tapatybę, šis modelis nėra saugus [CTW+20].

- Duomenų įvesties privatumas. Trečios šalys neturi matyti įvedamų duomenų. Tai gali būti tinklo saugumo spragos, duomenų surinkimo aplikacijų spragos ir t.t…

- Modelio išvesties privatumas. Modelio išvesties neturi matyti asmenys, kuriems šie duomenys nepriklauso. Šis punktas yra sąlyginis, priklauso nuo modelio svarbos. Jeigu tai yra svarbūs asmeniniai duomenys, negalima rizikuoti. Tačiau jeigu tai yra viešai prieinami duomenys, šis punktas negalioja.

- Modelio apsauga. Sukurtas modelis negali būti niekieno pasisavintas. Šis punktas yra skirtas apsaugoti programos kūrėją.

Tikslas ir uždaviniai

Darbo tikslas - ištirti ir palyginti privatumą saugančius dirbtinio intelekto algoritmus pagal jų saugumą, našumą ir panaudojamumą, bei pateikti rekomendacijas.

Darbo uždaviniai:

- Išanalizuoti esamus algoritmus pagal jų saugumą ir panaudojamumą.

- Identifikuoti kriterijus, kurių pagalba galima įvertinti privatumo išsaugojimą, bei palyginti algoritmus tarpusavyje.
- Ištirti kurie algoritmai yra realizuoti ir realizuoti algoritmus, kurie nėra atvirai prieinami.
- Palyginti algoritmus pagal našumą.

Tyrimo metodas

Algoritmai darbe bus tiriami pagal tikslo funkcijos kitimo greitį ir pagal duomenų atskleidimo rodiklį. Šis duomenų atskleidimo rodiklis bus skaičiuojamas pagal formulę (1).

$$DMDK = \sum_{n=0}^{m}(\sum_{k=0}^{h}(max((|\epsilon| + D_{eilut.:n,stulp.:k}) : \epsilon \in R))/h)/m \tag{1}$$

Prieš skaičiavimą reikia paimti visus modelio mokymui skirtus duomenis ir kiekvienai duomenų eilutei apskaičiuoti modelio išvestį. Skaičiavimus reikia atlikti tik su tomis eilutėmis, su kuriuomis modelis išvedė teisingą atsakymą. Turint tik tas eilutes, su kuriomis modelis išvedė teisingą atsakymą, galima į nelygybę įstatyti kintamuosius. Lygtyje yra naudojami tokie kintamieji: m - duomenų eilučių skaičius, h - parametrų skaičius (stulpeliai), ϵ - ieškomas didžiausias galimas kintamasis, su kuriuo modelis nepakeičia išvesties rezultatų, $D_{eilut.:n,stulp.:k}$ - duomenys n eilutėje ir k stulpelyje.

Laukiami rezultatai

Planuojamas darbo rezultatas - skirtingų algoritmų palyginimas ir rekomendacijos, kokius algoritmus reikia rinktis, ką daryti, kad pradiniai modelio duomenys būtų saugūs ir kriterijai, kurių pagalba galima įvertinti privatumo išsaugojimą.

Aktualūs šaltiniai

Darbe planuojama remtis įvairiais šaltiniais. Kalbant apie privatumo išsaugojimo temą, pagrindiniai šaltiniai yra:

- Privatumo išsaugojimas, taikant blokų-grandinių technologijas [ZZJ+20].
- Homomorfišio šifravimo sistema [Sen13].
- Saugumo problemos mašininiame mokymęsi [CLE+19].
- Straipsnis apie federuotą mašininį mokymąsi [Bha19].

Literatūros sąrašas

[116] Europos parlamento ir tarybos reglamentas, Apr 2016.

[Bha19] Dr. Santanu Bhattacharya. The new dawn of ai: Federated learning, Jan 2019.

[BJJ+18] Ho Bae, Jaehee Jang, Dahuin Jung, Hyemi Jang, Heonseok Ha, and Sungroh Yoon. Security and privacy issues in deep learning. *CoRR*, abs/1807.11655, 2018.

[CLE+19] Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks, 2019.

[CTW+20] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2020.

[Sen13] Jaydip Sen. Homomorphic encryption: Theory & applications, 07 2013.

[Tha20] Patricia Thaine. Perfectly privacy-preserving ai, Jan 2020.

[ZZJ+20] Yang Zhao, Jun Zhao, Linshan Jiang, Rui Tan, Dusit Niyato, Zengxiang Li, Lingjuan Lyu, and Yingbo Liu. Privacy-preserving blockchain-based federated learning for iot devices, 2020.