Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas

Paulius Milmantas

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Naugarduko g. 24, LT-03225 Vilnius
pauliusmilmantas@gmail.com

Santrauka. Straipsnyje yra atliekamas apmokymui skirtų duomenų saugumo tyrimas su skirtingais mašininio mokymosi modeliais. Modelių lyginimui apibrėžta metrika DMDK, kuri leidžia palyginti skirtingus modelius pagal jų pradinių mokymosi duomenų saugumo išsaugojimą. Maža DMDK reikšmė reiškia, kad tiriamas modelis yra linkęs atskleisti pradinius mokymosi duomenis ir nėra saugus. Atliktame tyrime pastebėta, kad „PyTorch neuroniniai tinklai" yra saugesni, nei homomorfiniu šifravimu grįstas „gradientinio nusileidimo modelis". Su visais analizuotais modeliais, išskyrus „PyTorch neuroninį tinklą", didėjant modelio tikslumui, didėja vidutinė DMDK reikšmė – modelis tampa saugesnis, o su „PyTorch neuroniniu tinklu", mažėja - modelis tampa mažiau saugus.

Raktiniai žodžiai: Mašininis mokymasis, duomenų saugumas, homomorfinis šifravimas.

1 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 iš duomenų, 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 [1].

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 [2]. Algoritmas apmoko modelį atpažinti duomenis. Turint sukurtą modelį, neturi būti galima atgaminti duomenų, pagal kuriuos jis buvo mokomas, bei negali būti identifikuo-

DOI: https://doi.org/10.15388/LMITT.2021.8

ti 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 [3].

- 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.

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

2 Tyrimo metodika

Siekiant palyginti, kaip modeliai gerai saugo privačius duomenis, apibrėžiame metriką DMDK (didžiausias modelio duomenų nuokrypis). Kuo metrika yra mažesnė, tuo tiksliau galima nuspėti, kokie duomenys buvo naudojami modelio mokymui. Ši metrika leidžia lyginti skirtingus modelius su skirtingais duomenimis.

DMDK metrika yra išvesta 1 lygtyje. Prieš DMDK skaičiavimą reikia paimti visus modelio mokymui skirtus duomenis ir kiekvienai duomenų eilutei apskaičiuoti modelio išvestį. Skaičiavimus reikia atlikti tik su tomis eilutėmis, su kuriomis 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.

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

1 lygtis. DMDK reikšmės skaičiavimas.

3 Tyrimo metodikos verifikavimas

Siekiant pademonstruoti, kad algoritmas teisingai įvertina skirtingus modelius ir galima juos palyginti tarpusavyje, paimkime kelis skirtingus modelių pavyzdžius

Tarkime, kad pirmas modelis turi viena parametrą – KMI. Pagal šį parametrą, modelis prognozuoja, ar žmogus serga cukriniu diabetu ar ne. Modelio tikslumas yra 54 %, jis visą laiką prognozuoja, kad žmogus serga cukriniu diabetu. Skaičiuojant DMDK, reikia pasirinkti tik tuos duomenis, su kuriais buvo išvesta teisinga prognozė. Modelis visą laiką prognozuoja, kad žmogus serga cukriniu diabetu, todėl, nėra tokios reikšmės, kurią pridėjus prie duomenų, pasikeis modelio prognozė. Todėl DMDK reikšmė yra ∞ ir tai reiškia, kad modelis negali būti saugesnis. Kai modelio prognozė visą laiką yra tapati, neįmano atgaminti pradinių duomenų, su kuriais modelis buvo mokomas.

Antras pavyzdinis modelis prognozuoja ar žmogus serga cukriniu diabetu, pagal 2 parametrus: KMI ir gimdymų skaičiumi. Modelio tikslumas yra 72 %, modelio DMDK reikšmė yra 0,00134. Sprendžiant pagal DMDK, modelis yra nesaugus ir yra labai priklausomas nuo pradinių duomenų, su kuriais modelis buvo mokomas. Modelio nesaugumui įrodyti, paimkime kelias savo sugalvotas duomenų eilutes ir pabandykime atgaminti pradinius duomenis. Tarkime, kad mūsų sugalvotos duomenų eilutės yra 1 lentelėje.

1 lentelė. Sugalvotos reikšmės modelio tikrinimui.

KMI	Gimdymų skaičius
25	3
25	2

Kiekvienai sugalvotai duomenų eilutei, analizuojame kiekvieną stulpelį. Prie stulpelio reikšmės pridedame ϵ ir didiname ϵ reikšmę tol, kol pasikeičia

modelio rezultatas. Analizuojame kiekvieną stulpelį ir pasirenkame stulpelį su mažiausia ϵ reikšme. Pakeičiame analizuojamos duomenų eilutės stulpelio reikšmę, su kuria gauname mažiausią ϵ reikšmę. Išanalizuokime pirmą duomenų eilutę 1 lentelėje. Po pirmos iteracijos, gauname pakeistą duomenų eilutę: KMI – 26, gimdymų skaičius – 3. Atliekame iteracijas tol, kol gaunamų naujų reikšmių skirtumas tampa labai mažas. Atlikus visų duomenų analizę, gauname duomenis, kurie buvo gauti iš sugalvotų duomenų eilučių. Jeigu modelio DMDK yra mažas, šie duomenys turi būti artimi pradiniams duomenims, kurie 12 buvo naudojami modelio mokymui. Atlikus analizę buvo gauti duomenys 2 lentelėje. Duomenys yra labai panašūs pradiniams modelio duomenims. Tai reiškia, kad DMDK matavimo algoritmas pasiteisino.

2 lentelė. Tikri modelio duomenys.

KMI	Gimdymų skaičius
26	1
22	2

Siekiant įsitikinti, kad algoritmas yra teisingas ir neduoda rezultatų, kurie tiesiogiai priklauso nuo duomenų, atlikti du eksperimentai.

1. Skaitant duomenis, jų visų reikšmės yra padalintos iš 10. Sukūrus modelį, apskaičiuota nauja DMDK reikšmė. Ji yra labai artima DMDK reikšmei, kai duomenys nebuvo dalijami iš 10. Tai reiškia, kad algoritmas nėra tiesiogiai priklausomas nuo duomenų, kuriais modelis yra mokomas.
2. Skaitant duomenis, prie jų reikšmių yra pridedamas atsitiktinai sugeneruotas triukšmas – skaičius nuo -1 iki 1. Naujo modelio DMDK reikšmė yra panaši į modelio, prieš duomenų keitimą. Tendencijos išlieka panašios.

4 Tyrimo rezultatai

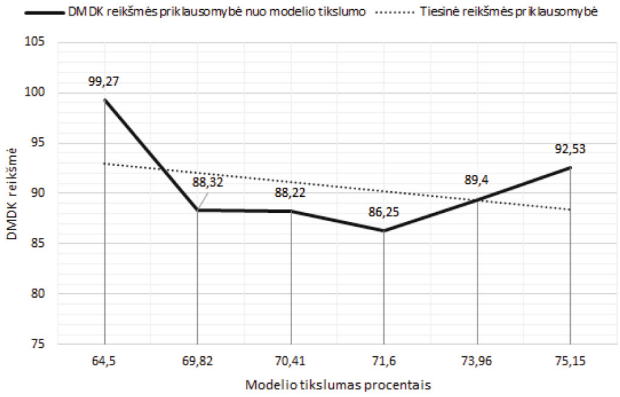
Tyrimo tikslams, buvo sukurti modeliai ir, skaičiuojant DMDK metriką, jie lyginami tarpusavyje. Tyrimui paimti diabetu sergančių žmonių duomenys iš „Sklear“ bibliotekos viešai prieinami duomenys. Gauti rezultatai pateikti 3 lentelėje.

3 lentelė. Tyrimo rezultatai.

Modelis	Nuostolių f-jos reikšmė (MSE)	Tikslumas	DMDK
„PyTorch neuroninis tinklas“	103.9837813	81.1 %	87.419
„Gradientinio nusileidimo algoritmas“	2272.06	63.78 %	0.051792
„Pallier homomorfiniu šifravimu grįstas modelis“	2185.37	63.52 %	0.1416

Pagal gautus tyrimo rezultatus, gauname tokias išvadas:

1. „PyTorch neuroninis tinklas“ greičiau artėja prie nuostolių funkcijos minimumo, nei kiti modeliai. Esant 70 % modelių tikslumui, „PyTorch modelis“ yra saugiausias. Mažiausiai saugus yra „gradientiniu nusileidimu grįstas modelis“.
2. „Pallier kriptografija grįstas modelis“ yra saugesnis už „gradientinio nusileidimo metodą“.
3. Didėjant „PyTorch modelio“ tikslumui, jis pradeda labiau prisirišti prie duomenų ir vidutinė galima maksimalaus nuokrypio reikšmė smarkiai krenta – modelis tampa vis mažiau saugus. Priklausomybė atvaizduota pav. 1. Jeigu uždavinio tikslas yra sukurti labai tikslų modelį, vertėtų apgalvoti ar „Pallier sistema grįstas modelis“ nėra geriau.
4. Su visais modeliais, išskyrus „PyTorch neuroninį tinklą“, didėjant modelio tikslumui, didėja vidutinė maksimalaus nuokrypio reikšmė – modelis tampa saugesnis, o su „PyTorch neuroniniu tinklu“, mažėja – modelis tampa mažiau saugus.



1 pav. DMDK reikšmės priklausomybė nuo „PyTorch modelio“ tikslumo

Tas pats tyrimas buvo atliktas su MNIST rašytinių skaitmenų nuotraukomis. Tik „PyTorch neuroninis tinklas“ su vienu paslėptu sluoksniu ir VGG16 konvergavo minimumo link. Pasiekus 74 % ir 82 % tikslumus, „PyTorch neuroninio tinklo“ DMDK reikšmės išlieka vienoda - 0.04999, MSE nuostolių funkcijos reikšmės atitinkamai 266.66730 ir 203.6602. VGG16 tinklui pasiekus 72 % tikslumą, MSE nuostolių funkcijos reikšmė - 2.59, DMDK reikšmė - 0.17. Reiškia, kad VGG16 modelis yra saugesnis už paprastą vieno paslėpto sluoksnio „PyTorch neuroninį tinklą“.

5 Išvados

Lyginant „PyTorch neuroninį tinklą“, „gradientinio nusileidimo algoritmą“ ir „Pallier homomorfiniu šifravimu grįstą modelį“, jeigu visi minėti modeliai vykdant tyrimą konverguoja minimumo link, pats saugiausias modelis yra „PyTorch neuroninis modelis“. Mažiausiai duomenis saugantis yra „gradientinio nusileidimo metodas“.

Naudojant „Pallier homomorfiniu šifravimu grįstą modelį“, egzistuoja koreliacija tarp modelio nuostolių funkcijos reikšmės ir DMDK rodiklio. Vidutiniškai, kuo nuostolių funkcijos reikšmė yra mažesnė, tuo DMDK rodiklis yra didesnis ir modelis yra saugesnis.

„Pallier homomorfiniu šifravimu grįstas modelis“, naudojantis „gradientinio nusileidimo algoritmą“, diverguoja su MNIST rašytinių skaičių nuotraukomis.

Paprastas „PyTorch neuroninis tinklas“ vienintelis iš minėtų modelių konverguoja minimumo link, su MNIST rašytinių skaičių nuotraukomis. Šiuo atveju, DMDK rodiklis vidutiniškai nukrenta iki 0.049 ir modelis tampa mažiau saugus, nei „Pallier algoritmu grįstas modelis“.

Literatūra

[1] 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.

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

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