Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas

Paulius Milmantas
Darbo vadovas: asist. dr. Linas Petkevičius

Vilniaus Universitetas
Matematikos ir informatikos fakultetas

Bakalauro darbo gynimas

Tiriama sritis

Mašininis mokymas yra dirbtinio intelekto sritis, kuri pasitelkia statistinius algoritmus, kad apibrėžtų duomenų generavimo mechanizmą, ar egzistuojančius sąryšius, priklausomybes.

Tikslas ir uždaviniai

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

Darbo tikslui įgyvendinti, iškelti šie **uždaviniai**:

1. Išanalizuoti esamus algoritmus pagal jų saugumą ir panaudojamumą.
2. Identifikuoti kriterijus, kurių pagalba galima įvertinti privatumo išsaugojimą, bei palyginti algoritmus tarpusavyje.
3. Ištirti kurie algoritmai yra realizuoti ir realizuoti dalį algoritmų, kurie nėra atvirai prieinami.
4. Palyginti algoritmus pagal našumą ir pateikti rekomendacijas.

Problematika

1. Turint sukurtą modelį, neturi būti galima atgaminti duomenų, pagal kuriuos jis buvo mokomas, bei negali būti identifikuoti asmenys [1].
2. Trečios šalys neturi matyti įvedamų duomenų. Tai gali būti tinklo saugumo spragos, duomenų surinkimo aplikacijų spragos ir t.t. . .
3. Modelio išvesties neturi matyti asmenys, kuriems šie duomenys nepriklauso.
4. Sukurtas modelis negali būti niekieno pasisavintas.

Modelių duomenų lyginimas (1)

$$atvirumas(s[r])_\theta = log_2|r| - log_2 rangas_\theta(s[r]) \quad (1)$$

Naudojama teorijoje, dėl sunkiai apskaičiuojamo rango [2].
s - duomenų rinkinys.
r $\in$ **R**, parenkamas atsitiktinai.

Modelių duomenų lyginimas (2)

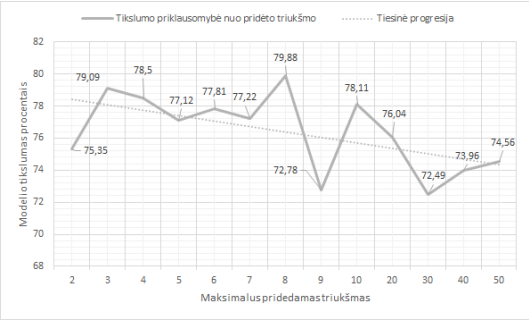
$$atvirumas(s[r])_\theta = -log_2 \int_0^{P_x(s[r])} \rho(x)dx \tag{2}$$

Dėl grafinės interpretacijos naudojama praktikoje.
Px - logaritminis entropijos matas.
s[**r**] - entropija yra $\rho(.)$ pasiskirstymo distribucijos.

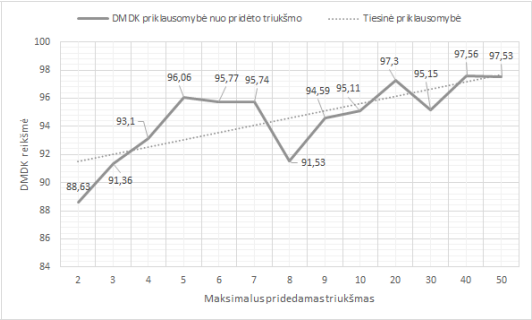
$$DMDK = \sum_{n=0}^{m}(\sum_{k=0}^{h}(max_{\epsilon}((|\epsilon| + D_{n,k}) : \epsilon \in R, modelis(|\epsilon| + D_{n,k}) = modelis(D_{n,k})))/h)/m \quad (3)$$

DMDK - Didžiausias modelio duomenų nuokrypis.
$\mathbf{D_{eilut:n,stulp:k}}$ - duomenys n eilutėje ir k stulpelyje.
ϵ - ieškomas didžiausias galimas kintamasis, su kuriuo modelis nepakeičia išvesties rezultatų.
m - duomenų eilučių skaičius.
h - parametrų skaičius (stulpeliai).

Pridėto triukšmo tyrimas

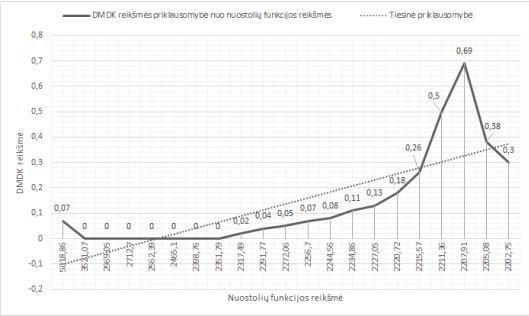


(a) Tikslumo tyrimas

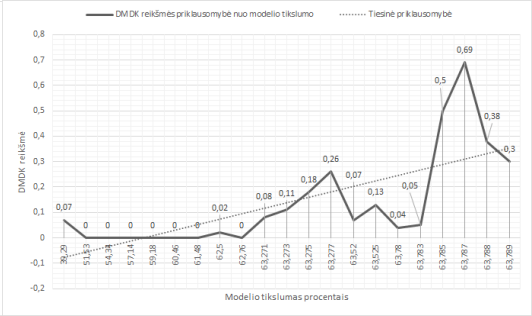


(b) DMDK tyrimas

Pallier tyrimas

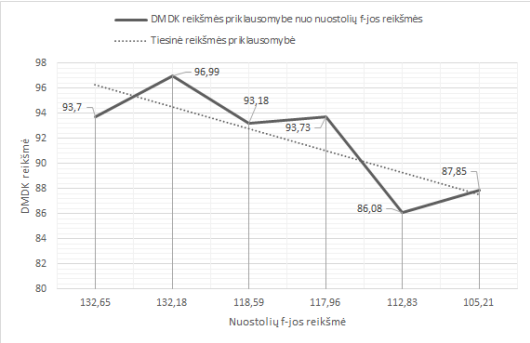


(c) DMDK tyrimas

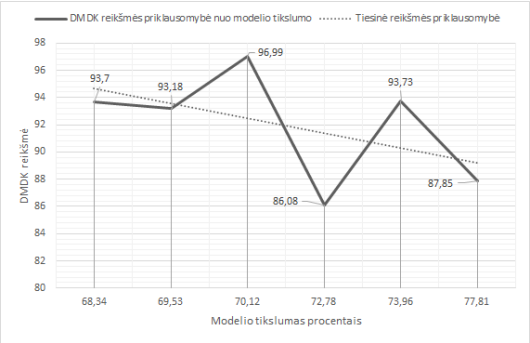


(d) Tikslumo tyrimas

PyTorch karkaso neuroninio tinklo tyrimas



(e) Nuostolių f-jos priklausomybės tyrimas



(f) Tikslumo tyrimas

Rezultatų aprobavimas

Baigiamojo darbo rezultatai buvo pristatyti nacionalinėje konferencijoje, o pranešimo tezės publikuotos konferencijos leidinyje:

Paulius Milmantas (2021) *Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas*, Vilnius University Open Series, pp. 71-76. doi: 10.15388/LMITT.2021.8.

Išvados

- Esant aukštam modelio tikslumui, rekomenduojama naudoti homomorfinį šifravimą. Esant mažesniam, nei 70% tikslumui, kai modelis priima < 20 parametrų, rekomenduojama naudoti PyTorch karkaso neuroninius tinklus.
- Esant didesniam modelio parametrų skaičiui, PyTorch karkaso neuroniniai tinklai labiau prisimena pradinius mokymosi duomenis ir juos galima lengviau atskleisti.
- Naudojant neuroninius tinklus be homomorfinio šifravimo ir modelio tikslumui esant daugiau nei 80%, rekomenduojama pridėti triukšmą prie pradinių modelio duomenų.
- Pradinių duomenų kiekis neturi įtakos modelio duomenų saugumui.

Šaltiniai

[1] Patricia Thaine. Perfectly privacy-preserving ai, 01 2020
[2] Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The secretsharer: Evaluating and testing unintended memorization in neural networks, 2019.