Vilniaus universitetas

Doktorantas:
Paulius Vaitkevičius

Vadovas:
Dr. Virginijus Marcinkevičius

Pirmųjų metų ataskaita
2019 m. spalio 31 d.

Mašininiu mokymusi grįstų atvirųjų šaltinių žvalgybos informacijos išskyrimo ir analizės metodai

Vilniaus universitetas

TURINYS

1. Problemos apibrėžimas, tyrimo objektas, tikslai ir planuojami gauti rezultatai
2. Ataskaitinių metų darbo planas ir ataskaita
3. Kitų metų darbo planas
4. Trumpas per metus gautų mokslinių rezultatų pristatymas

PROBLEMOS APIBRĖŽIMAS, TYRIMO OBJEKTAS, TIKSLAI IR PLANUOJAMI GAUTI REZULTATAI

Vilniaus universitetas

Problemos apibrėžimas

ESAMA SITUACIJA

- Egzistuojantys duomenų išviliojimo tinklapių *(angl. Phishing)* atpažinimo algoritmai neatsparūs:
  - besikeičiančiai aplinkai (pvz. URL formatai, SSL sertifikatai, kt.),
  - apsimetinėjimo atakoms *(angl. Adversarial Learning)*, kai bandoma apmokyti ML algoritmą, pasinaudoti kito ML algoritmo mokymosi spragomis.

SIEKIAMA SITUACIJA

- Efektyvus duomenų išviliojimo tinklapių atpažinimo algoritmas su tokiomis savybėmis:
  - atsparus aplinkos pokyčiams - prisitaikantis per automatinį klasifikavimo požymių parinkimą ir nuolatinį mokymąsi,
  - atsparus apsimetinėjimo atakoms.

Vilniaus universitetas

Tyrimo tikslas

Sukurti apsimetinėjimo atakoms atsparų metodą, grįstą giliaisiais neuroniniais tinklais ir natūralios kalbos apdorojimo algoritmais, kuris leistų efektyviai ir patikimai atpažinti duomenų išviliojimo internete tinklapius.

Vilniaus universitetas

Tyrimo objektas

1. Mašininio mokymo ir giliojo mašininio mokymo algoritmai, skirti atpažinti duomenų išviliojimo internete (angl. „Phishing“) tinklapius.
2. Atsparūs priešiškoms atakoms algoritmai (angl. „Adversarial Machine Learning“).

Vilniaus universitetas

Tyrimo uždaviniai

1. Atlikti literatūros analizę, išanalizuoti state-of-the-art algoritmus duomenų išviliojimo internete tinklapių atpažinimui.
2. Atkartoti *state-of-the-art* algoritmų rezultatus.
3. Pasiūlyti naują efektyvesnį duomenų išviliojimo internete tinklapių atpažinimo metodą.
4. Sukurti duomenų rinkinius eksperimentų vykdymui.
5. Atlikti eksperimentinius tyrimus, palyginant pasiūlytą metodą su *state-of-the-art* algoritmais.

Vilniaus universitetas

Planuojami rezultatai

1. Atlikta literatūros analizė, palyginant pažangiausius algoritmus, naudojamus asmens duomenų išviliojimo internete tinklapiams atpažinti;
2. Eksperimentiniai tyrimai:
   a. Mašininio mokymosi algoritmų efektyvumo palyginimas, panaudojant viešai prieinamus sukčiavimo internete tinklapių URL duomenų rinkinius, su iš anksto išskirtais požymiais.
   b. Giliojo mašininio mokymosi algoritmų efektyvumo palyginimas, panaudojant viešai prieinamus sukčiavimo internete tinklapių URL duomenų rinkinius, su iš anksto išskirtais požymiais.
   c. Giliojo mašininio mokymosi algoritmų (RNN, LSTM, kt.) efektyvumo tyrimai, naudojant natūralaus teksto apdorojimo technikas (N-grams, word embedings, kt.).
   d. Įgyvendintų giliojo mašininio mokymosi algoritmų modifikacijos, ar naujų algoritmų kūrimas, sprendžiant apibrėžtus uždavinius.
   e. Pasiūlyto giliojo mašininio mokymo algoritmo eksperimentinis tyrimas analizuojant jo efektyvumą;
   f. Pasiūlyto giliojo mašininio mokymo algoritmo atsparumo priešiškoms atakoms (angl. „Adversarial Machine Learning) eksperimentiniai tyrimai.

ATASKAITINIŲ METŲ DARBO PLANAS IR ATASKAITA

KITŲ METŲ DARBO PLANAS

Vilniaus universitetas

Rezultatai ir terminai	Komentarai	SM-I		SM-II		SM-III		SM-IV	
		S-1	S-2	S-3	S-4	S-5	S-6	S-7	S-8
DALYVAVIMAS KONFERENCIJOSE ir KT.									
1. Dalyvavimas konferencijoje Lietuvoje.	KODI-2019 / 2019-10-04								
2. Dalyvavimas konferencijoje Lietuvoje.	DAMSS-2019 / 2019-11-29								
3. Tyrimo rezultatų pristatymas tarptautinėje mokslinėje konferencijoje.	DL-2019 / 2019-07-26, Varšuva								
4. Tyrimo rezultatų pristatymas tarptautinėje mokslinėje konferencijoje.	Konferencija artimajame užsienyje								
PLANUOJAMAS MOKSLINIŲ TYRIMŲ PUBLIKAVIMAS									
1. Mokslinių tyrimų disertacijos tema apžvalga (konferencijos darbų medžiagoje)	DAMSS-2019								
2. Teorinio tyrimo publikavimas (recenzuojamoje konferencijos darbų medžiagoje)	Konferencija artimajame užsienyje								
3. Algoritmų palyginimo tyrimo rezultatų publikavimas (recenzuojamame leidinyje)	INFORMATICA								
4. Empirinio tyrimo rezultatų publikavimas (recenzuojamame leidinyje)									
STUDIJOS									
1. Informatikos ir informatikos inžinerijos tyrimo metodai ir metodika			10						
2. Fundamentalieji informatikos ir informatikos inžinerijos metodai									
3. Didžiųjų duomenų analitika									
4. Mašininis mokymasis									
X. Bendruosius gebėjimus stiprinančios veiklos (3 kreditai)	1,95 kredito + DL-2019								
MOKSLINIŲ TYRIMŲ IR DISERTACIJOS RENGIMAS									
1. Mokslinių tyrimų disertacijos tema apžvalga ir analizė									
2. Mokslinio tyrimo vykdymas:									
2.1. Tyrimo metodikos sudarymas									
2.2. Teorinis tyrimas									
2.3. Empirinis tyrimas									
3. Atskirų daktaro disertacijos dalių parengimas									
4. Daktaro disertacijos parengimas ir svarstymas padalinyje									
5. Daktaro disertacijos gynimas									

ataskaitiniai metai

kiti metai

Vilniaus universitetas

Kitų metų darbo planas (detalus)

1. **Moksliniai tyrimai**
   a. Learning Phishing Websites URLs Using Long Short-Term Memory Network and Gated Recurrent Unit (DAMSS)
   b. Sukčiavimo internete atpažinimo (angl. „Phishing“) metodo su automatiniu požymių atpažinimu kūrimas, giliųjų neuroninių tinklų (angl. „Deep Neural Networks“,) ir natūralios kalbos apdorojimo (angl. „Natural Language Processing, NLP“) pagalba.
2. **Disertacijos rengimo etapai:**
   a. Teorinis tyrimas;
   b. Empirinis tyrimas
3. **Dalyvavimas konferencijose**
   a. Dalyvavimas konferencijoje Lietuvoje (DAMSS);
   b. Tyrimo rezultatų pristatymas tarptautinėje mokslinėje konferencijoje.
4. **Publikacijų rengimas**
   a. Mokslinių tyrimų disertacijos tema apžvalga (konferencijos darbų medžiagoje) (DAMSS)
   b. Teorinio tyrimo publikavimas (recenzuojamoje konferencijos darbų medžiagoje).

TRUMPAS PER METUS GAUTŲ MOKSLINIŲ REZULTATŲ PRISTATYMAS

Tyrimas Nr. 1

Klasikinių klasifikavimo algoritmų palyginimas asmens duomenų išviliojimo tinklalapiams aptikti

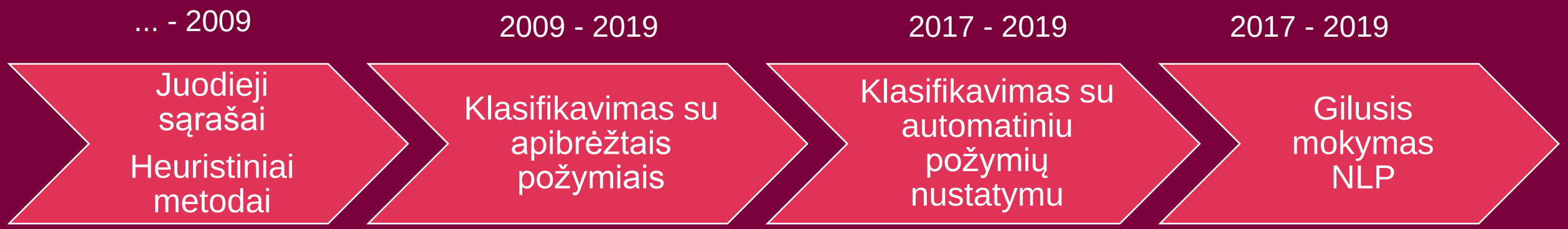
Vilniaus universitetas

Tyrimas

1. Literatūros apžvalga
2. Tyrimo klausimas
3. Priemonės: duomenų rinkiniai ir algoritmai
4. Tyrimo metodas
5. Rezultatai
6. Išvados
7. Tolimesni tyrimai

Vilniaus universitetas

Literatūros apžvalga



Vilniaus universitetas

Literatūros apžvalga

Year	Authors	Classifier	Dataset # phish.	Dataset # legit.	Accuracy
2017	Marchal et al. [18]	Gradient Boosting	100,000	1000	99.90%
2010	Whittaker et al. [39]	Logistic Regression	16,967	1,499,109	99.90%
2011	Xiang et al. [41]	Bayesian Network	8,118	4,780	99.60%
2018	Cui et al. [8]	C4.5	24,520	138,925	99.78%
2013	Zhao et al. [44]	Classic Perceptron	990,000	10,000	99.49%
2018	Patil et al. [22]	Random Forest	26,041	26,041	99.44%
2013	Zhao et al. [44]	Label Efficient Perceptron	990,000	10,000	99.41%
2014	Chen et al. [6]	Logistic Regression	1,945	404	99.40%
2018	Cui et al. [8]	SVM	24,520	138,925	99.39%
2018	Patil et al. [22]	Fast Decision Tree Learner (REPTree)	26,041	26,041	99.19%

Image source: author

Vilniaus universitetas

Tyrimo klausimas

„Kurie klasikiniai klasifikavimo algoritmai labiausiai tinka fišingo tinklapių aptikimui, nepriklausomai nuo konkretaus fišingo URL duomenų rinkinio?“

10
klasifikatorių

3
duomenų
rinkiniai

Image source:
lynda.com

Vilniaus universitetas

Naudoti duomenų rinkiniai

1. Duomenų rinkinys iš University of California Irvine (UCI) repository, contributed by Mohammad et al.
   [30 požymių, 11.055 įrašų]
2. Duomenų rinkinys iš Mendeley Datasets, contributed by Choon Lin Tan
   [48 požymių, 10.000 įrašų]
3. Duomenų rinkinys iš UCI repository, contributed by Abdelhamid et al.
   [8 požymių, 1.353 įrašų]

```
@attribute NumDots numeric
@attribute SubdomainLevel numeric
@attribute PathLevel numeric
@attribute UrlLength numeric
@attribute NumDash numeric
@attribute NumDashInHostname numeric
@attribute AtSymbol numeric
@attribute TildeSymbol numeric
@attribute NumUnderscore numeric
@attribute NumPercent numeric
@attribute NumQueryComponents numeric
@attribute NumAmpersand numeric
@attribute NumHash numeric
@attribute NumNumericChars numeric
@attribute NoHttps numeric
@attribute RandomString numeric
@attribute IpAddress numeric
@attribute DomainInSubdomains numeric
@attribute DomainInPaths numeric
@attribute HttpsInHostname numeric
@attribute HostnameLength numeric
@attribute PathLength numeric
```

Image source: author

Vilniaus universitetas

Naudoti klasifikatoriai

1. AdaBoost
2. Classification and Regression Tree (CART)
3. Gradient Tree Boosting
4. k-Nearest Neighbors
5. Multilayer Perceptron (MLP) with backpropagation
6. Naïve-Bayes
7. Random Forest
8. Support-vector Machine (SVM) with linear kernel
9. Support-vector Machine with 1st degree polynomial kernel
10. Support-vector Machine with 2nd degree polynomial

Vilniaus universitetas

Tyrimo metodas (1)

1. Kiekvienam duomenų rinkiniui (3):
    1.1 Kiekvienam klasifikavimo algoritmui (10):
        1.1.1 Sukonfigūruoti algoritmą *(Python 3.7 programavimo kalba, Scikit-Learn biblioteka)*
        1.1.2 Parinkti geriausius *hiper-parametrus*:
            1.1.2.1 Parinkti hiper-parametrus
            1.1.2.2 Apmokyti ir ištestuoti algoritmą naudojant kryžminę validaciją *(angl. cross-validation),* su 30 imčių *(angl. folds),* vertinant pagal klasifikavimo tikslumą (angl. classification accuracy)
            1.1.2.3 Nupiešti mokymosi kreives ir įvertinti:
                - ar algoritmas mokosi tinkamai
                - ar algoritmas persimoko *(angl. overfitting)*
                - ar algoritmas nesimoko *(angl. underfitting)*
            1.1.2.4 Jei reikalinga, kartoti nuo žingsnio 1.1.2.1 su pagerintais hiper-parametrais

Vilniaus universitetas

Tyrimo metodas (2)

- Jei atotrūkis didelis arba mokymosi kreivė plokščia ties 1: **algoritmas persimoko**
  - Sumažinti požymių skaičių
  - Padidinti mokymo aibės dydį
  - Sumažinti algoritmo kompleksiškumą, pvz.:
    - Sumažinti medžių/šakų skaičių ir (arba) gylį
    - Sumažinti neuroninio tinklo sluoksnių skaičių
  - Padidinti reguliarizacijos parametrą
- Jei atotrūkis mažas: **algoritmas nesimoko**
  - Padidinti požymių skaičių, naudoti polinominius požymius
  - Padidinti algoritmo kompleksiškumą, pvz.:
    - Padidinti medžių/šakų skaičių ir (arba) gylį
    - Padidinti neuroninio tinklo sluoksnių skaičių
  - Sumažinti reguliarizacijos parametrą

Image source: author

Vilniaus universitetas

Tyrimo metodas (3)

1. Kiekvienam duomenų rinkiniui (3):
   1.1 Kiekvienam klasifikavimo algoritmui (10):
      ...
      1.1.3 Patikrinti klasifikavimo tikslumų aibių normalumo prielaidas:
         - Shapiro-Wilk testas,
         - D'Agostino's K^2 testas,
         - Anderson-Darling testas,
         - Histograma,
         - Q-Q grafikas

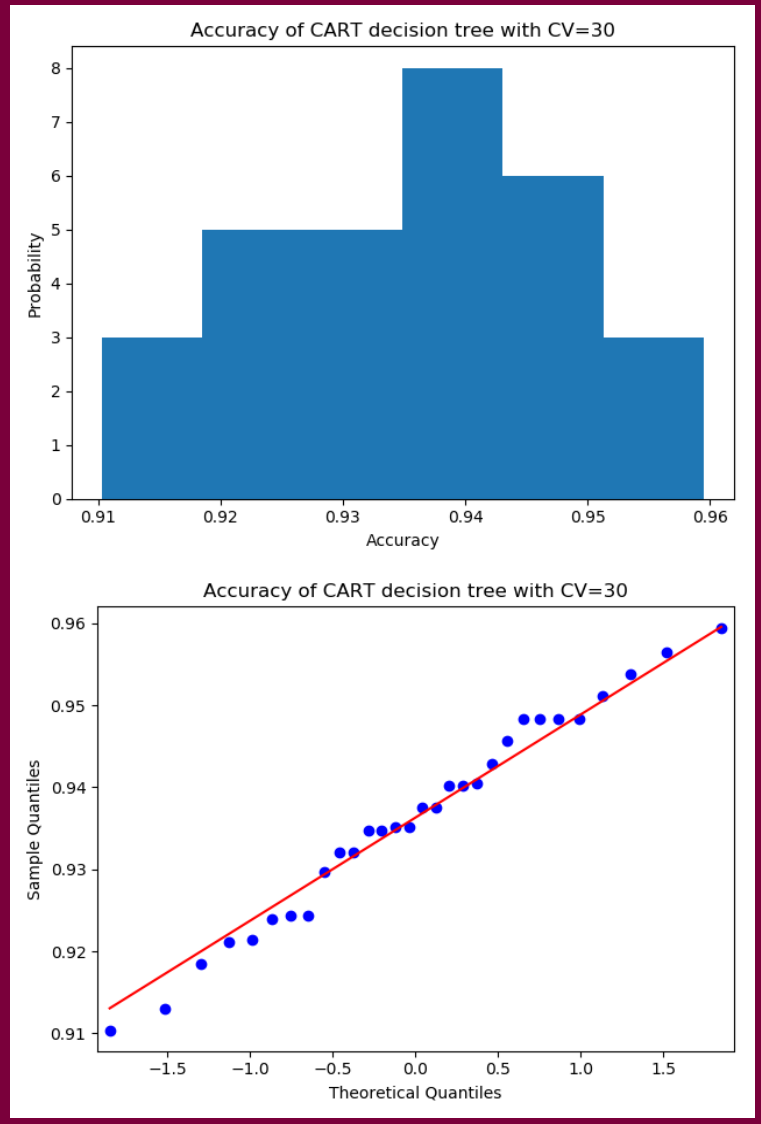


Image source: author

Vilniaus universitetas

Tyrimo metodas (4)

2. Kiekvienam duomenų rinkiniui (3):
   ...
   2.2 Patikrinti klasifikavimo tikslumo skirtumų statistinį reikšmingumą:
   - Student‘s T-test
   - Reikšmingumo kriterijus $\alpha = 0.05$
   - Jei *p-value* > α – negalima atmesti H_0
   - Tikrinamos hipotezės:
     - H_0: dvi testuojamos populiacijos neturi statistiškai reikšmingo skirtumo
     - H_A: egzistuoja statistiškai reikšmingas skirtumas tarp testuojamų populiacijų

Algorithm	UCI-2015	UCI-2016	MDP-2018
AdaBoost	0,9352	0,8495	0,9728
CART	0,9363	0,8930	0,9574
Gradient Tree Boosting	0,9381	**0,9034**	**0,9742**
k-Nearest Neighbors	0,9481	0,8641	0,8564
Naïve-Bayes	0,9057	0,8225	0,9177
Multilayer Perceptron	**0,9722**	0,9028	0,9671
Random Forest	0,9525	0,8916	0,9715
SVM with linear kernel	0,9271	0,8365	0,9422
SVM with 1st deg. pol. kernel	0,9257	0,8328	0,9334
SVM with 2nd deg. pol. kernel	0,9388	0,7152	0,9549

Image source: author

Vilniaus universitetas

Tyrimo metodas (6)

1. Kiekvienam duomenų rinkiniui (3):
   ...
   1.3 Suranguoti rezultatus, atsižvelgiant į klasifikavimo tikslumo skirtumų statistinį reikšmingumą:
   - *Standard competition ranking ("1-2-2-4")*
   - *Fractional ranking ("1-2.5-2.5-4")*
   - *Dense ranking ("1-2-2-3")*
2. Apjungti rangavimo rezultatus į bendrą reitingą

Algorithm	SCR points	FR points	DR points
Multilayer Perceptron	27	25.5	29
Gradient Tree Boosting	29	24	29
Random Forest	29	24	29
AdaBoost	25	19.5	28
CART	25	20.5	27
SVM with 2nd deg. pol. kernel	16	13	25
k-Nearest Neighbors	16	11.5	23
SVM with linear kernel	13	10	24
SVM with 1st deg. pol. kernel	13	10	24
Naïve-Bayes	9	7	22

Image source: author

Vilniaus universitetas

Rezultatai, išvados

Eksperimento metu nustatyta, kad:

1. Iškeltam uždaviniui spręsti geriausiai tinka daugiasluoksnis perceptronas ir ansamblio tipo klasifikatoriai.
2. Panašumu paremti *(angl. instance similarity)* ir tikimybiniai klasifikatoriai uždaviniui spręsti tinka mažiausiai.
3. Eksperimento rezultatai iš esmės sutampa su literatūros tyrimo metu nustatytomis tendencijomis.
4. Literatūros apžvalgoje nustatyti ant nesubalansuotų duomenų aibių pirmaujantys klasifikatoriai (CART, SVM, kt.), eksperimento metu neparodė tokių aukštų rezultatų.

Algorithm	SCR points	FR points	DR points
Multilayer Perceptron	27	25.5	29
Gradient Tree Boosting	29	24	29
Random Forest	29	24	29
AdaBoost	25	19.5	28
CART	25	20.5	27
SVM with 2nd deg. pol. kernel	16	13	25
k-Nearest Neighbors	16	11.5	23
SVM with linear kernel	13	10	24
SVM with 1st deg. pol. kernel	13	10	24
Naïve-Bayes	9	7	22

Image source: author

Tyrimas Nr. 2

Giliojo mašininio mokymosi algoritmų efektyvumo palyginimas, panaudojant viešai prieinamus sukčiavimo internete tinklapių URL duomenų rinkinius, su iš anksto išskirtais požymiais

Vilniaus universitetas

Konfigūracija ir rezultatai

Pirmieji bandymai su DNN:

- Duomenų aibės padalijimas:
  - TRAIN: 80%;
  - CV: 10%;
  - TEST: 10%
- Iteracijos: 10
- Rezultatai:
  - **97.07%** vid. tikslumas (accuracy) ant Mohammad et al. duomenų aibės (97.22% ant MLP)
  - **97.52%** vid. tikslumas (accuracy) ant Choon Lin Tan duomenų aibės (97.42% on GTB)

TensorFlow 2.0 Alpha

- **Model**: keras.Sequential()
- **Hidden layers:** 3
- **Hidden layer size**: 300 – 500
- **Activation**: ReLU, Softmax (output)
- **Optimizer**: Adam
- **Loss function**: Sparse Categorical Cross-entropy
- **CV**: Early stopping, patience = 2

Vilniaus universitetas

Tolimesni tyrimai

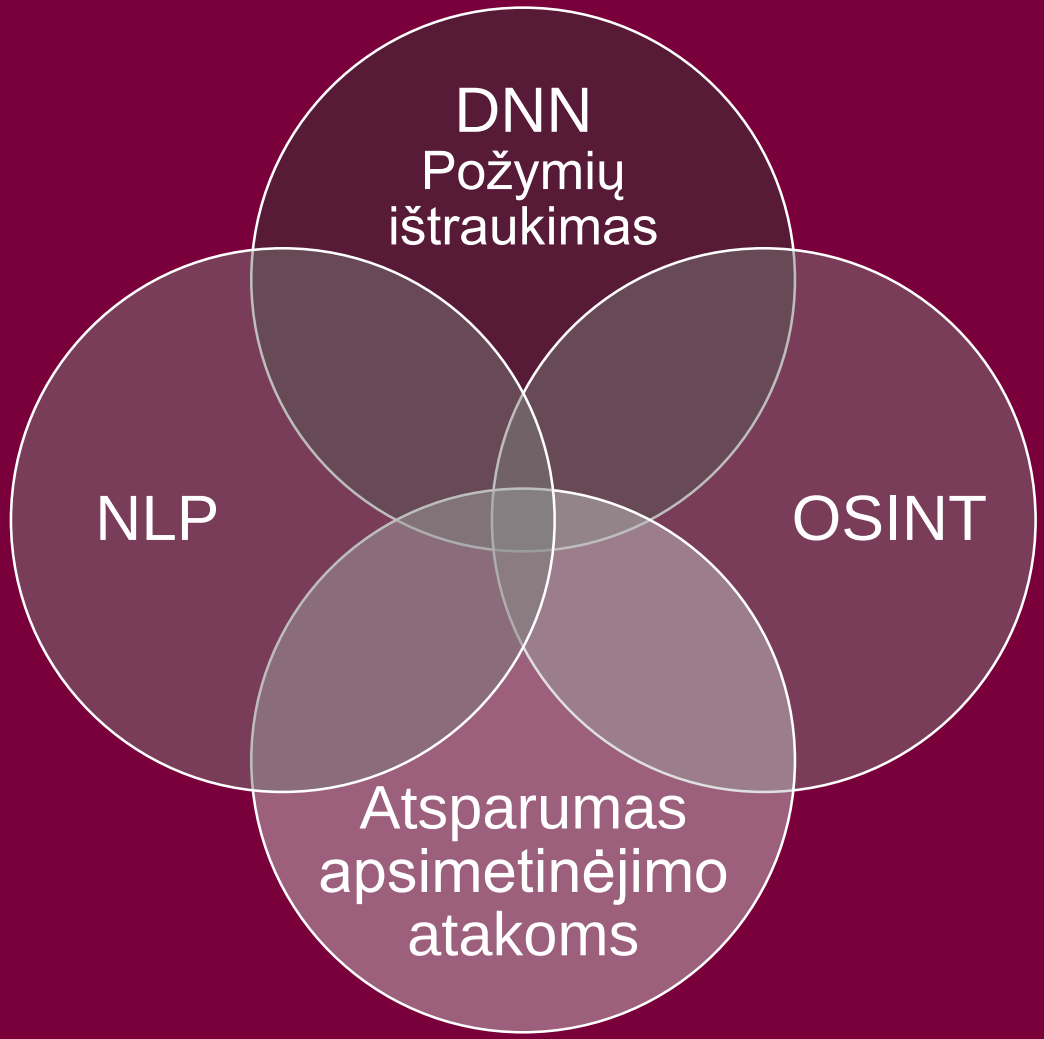


Image source: author

Vilniaus universitetas

AČIŪ UŽ DĖMESĮ

ORCID

Paulius Vaitkevičius
VU DMSTI doktorantas
+370 650 83623
paulius.vaitkevicius@mif.vu.lt