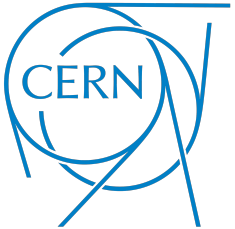
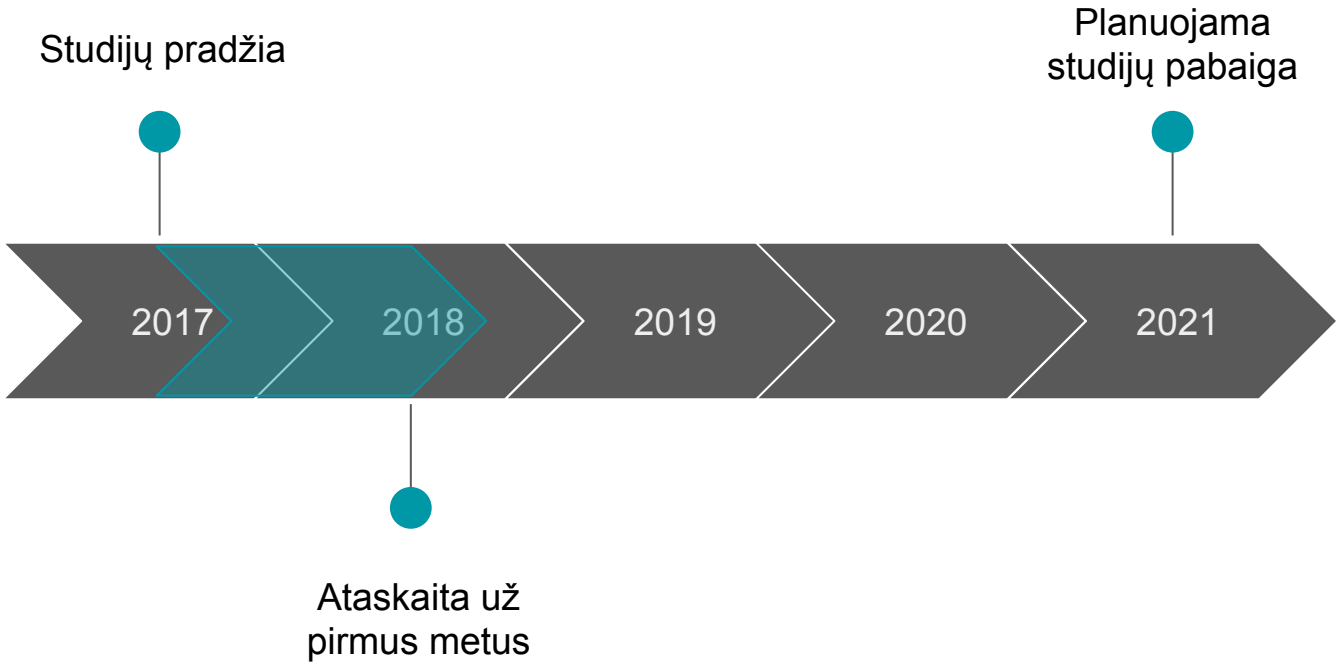


Europos branduolinių mokslinių tyrimų organizacijos (CERN) kompaktiško miuonų solenoido (CMS) eksperimento duomenų sertifikavimas pasitelkiant mašininio mokymo metodus

Doktorantas: Mantas Stankevičius
Vadovas: Virginijus Marcinkevičius
Konsultantas: Valdas Rapševičius

Studijos


Studijų pradžia
Planuojama studijų pabaiga
2017
2018
2019
2020
2021
Ataskaita už pirmus metus

Tikslas

Sukurti klasifikavimo metodus skirtus CERN CMS duomenų sertifikavimui naudojant mašininį mokymąsi

Objektas

- CERN CMS duomenys
- Klasifikavimo algoritmai
- Anomalijų paieškos algoritmai
- Mašininis mokymasis

Ataskaitinių metų planas

- ✔ Išlaikyti 2 dalykų egzaminus
- ✔ Publikuoti straipsnį
- ✔ Dalyvauti tarptautinėje konferencijoje

Ataskaitinių metų rezultatai

Puikiai išlaikyti 2 egzaminai:

- Duomenų analizės strategijos ir sprendimų priėmimas
- Lygiagretieji ir paskirstytieji skaičiavimai

Ataskaitinių metų rezultatai

Atliktas esamų mašininio mokymo metodų, naudojamų fizikinių duomenų sertifikavimui, tyrimas ir pritaikymas

Metodų tyrimo apibendrinimas pateiktas publikuotame straipsnyje:

Comparison of Supervised Machine Learning Techniques for CERN CMS Offline Data Certification. Joint Proceedings of Baltic DB&IS 2018 Conference Forum and Doctoral Consortium, Trakai, Lithuania, July 1-4, 2018. CEUR-WS.org, online http://ceur-ws.org/Vol-2158/paper18dc6.pdf (ISSN 1613-0073)

Ataskaitinių metų rezultatai

Skaitytas pranešimas tarptautinėje konferencijoje *Baltic DB&IS 2018*

Comparison of Supervised Machine Learning Techniques for CERN CMS Offline Data Certification.

Atlikti tyrimai ir pasiekti rezultatai

Data Quality Monitoring

Neapdorotų duomenų ir detektoriaus statuso srautas nuolat yra rašomas į duomenų saugyklą

Maža dalis duomenų yra rekonstruojama ir stebima realiu laiku.

Online shifters pažymi Run kaip GERAS jeigu yra pakankamai naudingų duomenų ir įranga veikė puikiai

Duomenys yra pilnai rekonstruojami ir kalibruojami po maždaug 48 valandų

Offline shifters patikrina svarbiausias histogramas ir nustato duomenų gerumą

Sertifikavimas yra daromas Run ir Lumisekcijų lygyje

GoldenJSON yra sudaromas tik iš gerų Run ir Lumisekcijų

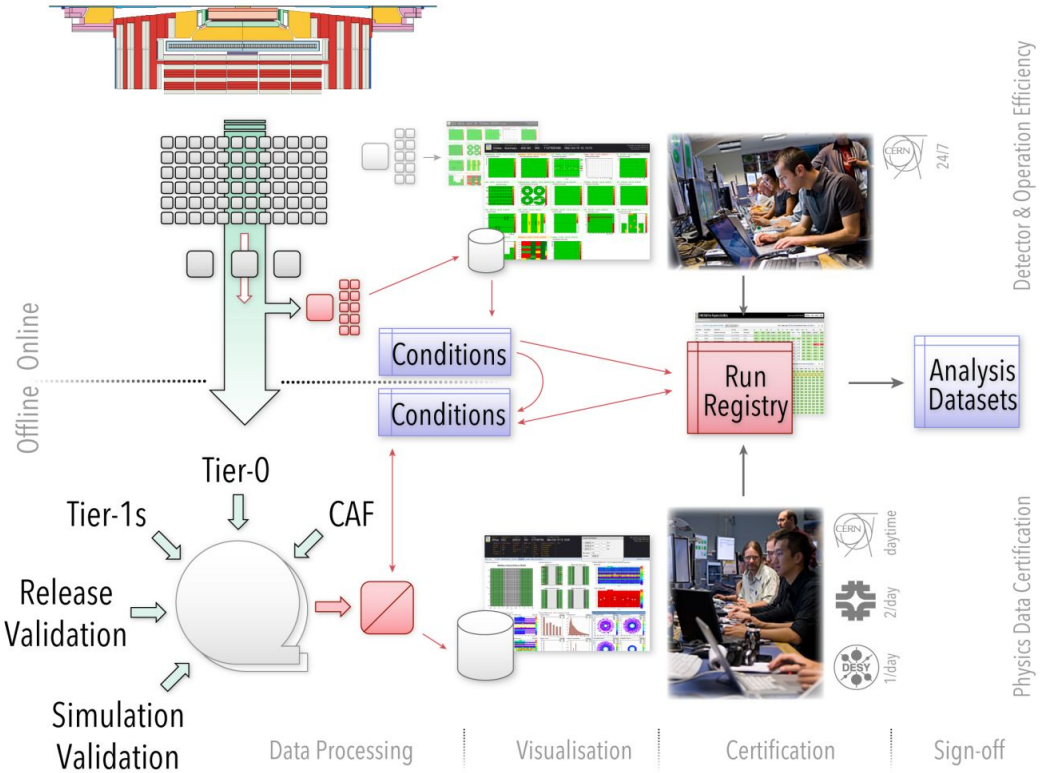


Image source http://iopscience.iop.org/article/10.1088/1742-6596/664/7/072039/meta

Uždavinys

Palyginti skirtingus dviklasio klasifikavimo metodus ir nustatyti tinkamiausią skirtą CERN CMS duomenų sertifikavimui

Duomenys

Naudojami CERN CMS eksperimento 2016 metų duomenys

- 160.000 lumisekcijų
- 2807 požymiai
- 2 išbalansuotos klasės (49:1)

Paruošimas:

- Požymių normalizavimas
- Klasių reikšmės imamos iš GoldenJSON

Įranga

Programinė įranga:

- Python (v3.6)
- Keras (v2.1.5)
- Tensorflow (v1.7)
- XGBoost (v0.71)
- scikit-learn (v0.19.1)

Kompiuterinė įranga:

- PC + NVIDIA GPU (GeForce GTX 1080 Ti)
- VM (8 cores 2.2 GHz, 16GB RAM)

Klasifikavimo metodai

Support Vector Machine (SVM)

- Scikit-learn biblioteka
- Didelis kiekis požymių neigiamai įtakojo greitaveiką
- Metodas pašalintas iš tolimesnių tyrimų

Gaussian Naive Bayes (NB)

- Scikit-learn biblioteka
- Be parametrų

- Greitas apmokymas
- Prastas tikslumas

Klasifikavimo metodai: Medžiai

Random Forest (RF)

- Scikit-learn biblioteka
- 64 medžiai su 7 lygių gyliu

- Greitas apmokymas
- Geras tikslumas
- Didelis miškas lėčiau klasifikuoja

Gradient Boosted trees (XGB)

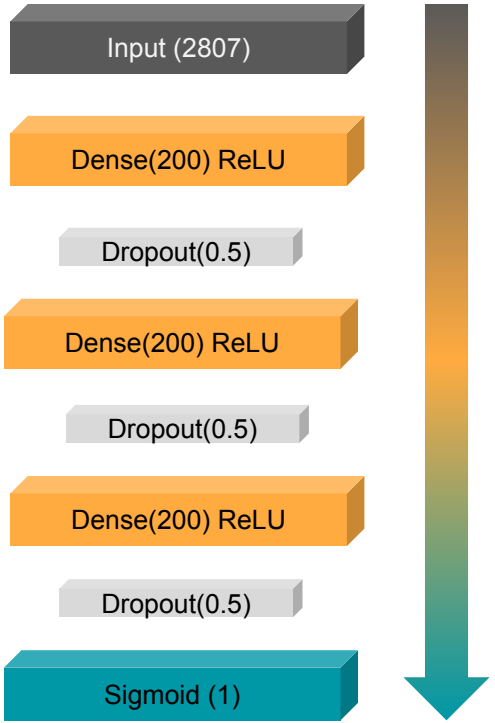
- XGBoost biblioteka
- 64 medžiai su 7 lygių gyliu

- Geras tikslumas
- Vidutiniškas apmokymo laikas
- Didelės atminties sąnaudos apmokymo metu

Klasifikavimo metodai: Neuroninis tinklas

- *Keras* biblioteka su *Tensorflow-GPU* backend
- 3 paslėpti sluoksniai naudoja *ReLU* aktyvacijos funkciją
- Kiekvieną paslėptą sluoksnį seka *Dropout* sluoksnis
- Išėjimo sluoksnis naudoja *Sigmoid* aktyvacijos funkciją
- *Early stopping* pradeda išvengti permokymo

- Vidutiniškas tikslumas
- Labai lėta hyper parametrų paieška

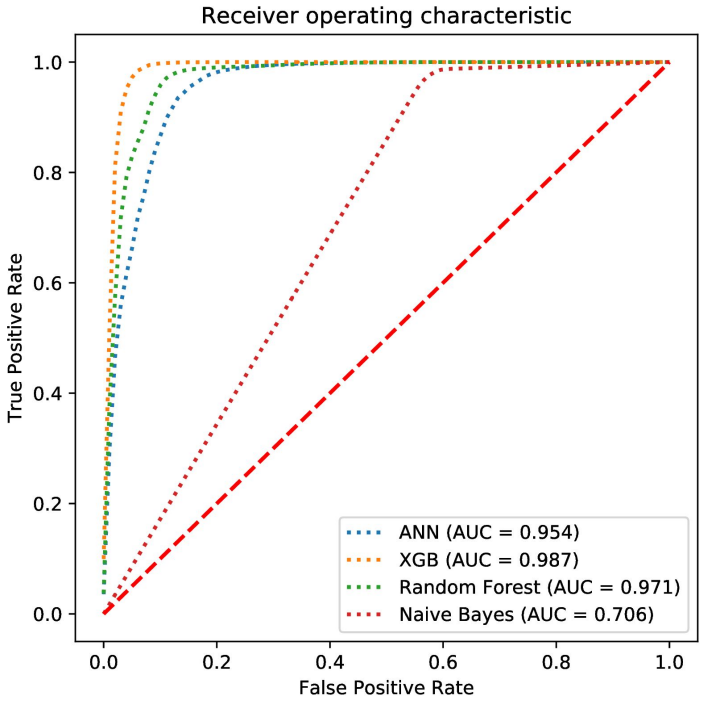


Eksperimentai: #1

Validacija:
Shuffle stratified 10 fold cross validation

Metrikos:
AUC, ACC, F_1 score

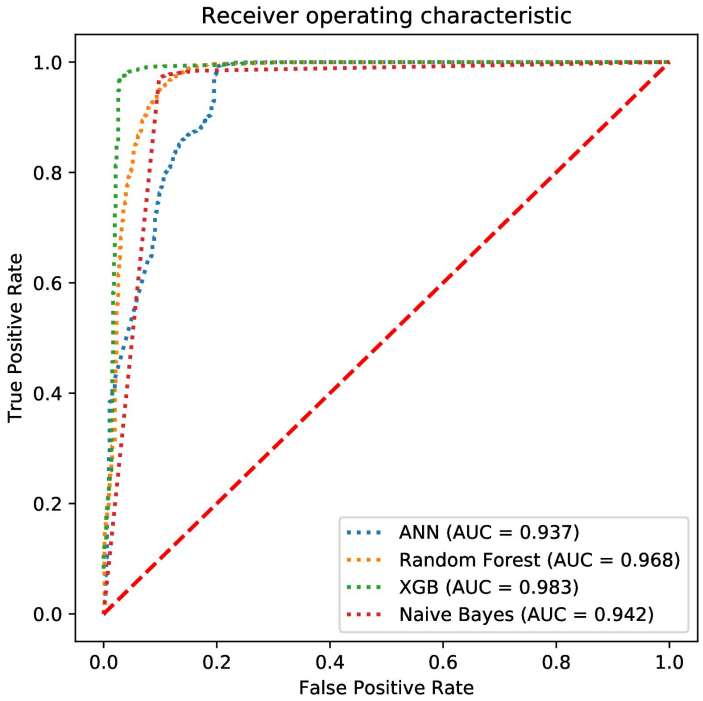
	AUC	ACC	F_1
XGB	**0.987**	**0.997**	**0.998**
Random Forest	0.970	0.980	0.990
ANN	0.954	0.961	0.79
Naive Bayes	0.706	0.971	0.985
		Vidurkiai	



Eksperimentai: #2

- Duomenys surūšiuojami pagal laiką
- Apmokymui - senesni duomenimis
- Klasifikavimui - naujesni duomenys

Naive Bayes metodo rezultatai 25% geresni nei pirmame eksperimente



Išvados

Išbandžius 5 populiariausius *supervised* mokymosi metodus, gautos šios išvados:

- Didžiausias vidutinis ROC AUC įvertis 0.987 gaunamas naudojant Gradient Boosted Trees metodą (XGBoost biblioteka)
- Prasčiausi rezultatai gauti naudojant tikimybinį Naive Bayes metodą. Vidutinis ROC AUC įvertis 0.7

Kitų metų planas

- Išlaikyti 2 egzaminus
- Anomalijų CERN CMS duomenyse analizė
- Mašininio mokymo ir gilaus mokymo metodų naudojamų anomalijų CERN CMS duomenyse aptikimui analizė ir taikymas
- Gautus rezultatus publikuoti recenzuojamame žurnale
- Dalyvauti tarptautinėje konferencijoje

Ačiū

Už dėmesį