Algoritmi strojnog učenja za klasifikaciju fragmenata dodatoteka (Petra Omrčen)

Zadatak

U okviru seminara treba istražiti koji algoritmi strojnog učenja su primjenjivi za problem klasifikacije fragmenata datoteka. Fragmenti datoteka su zapisi duljine 512 bajtova koji su isječci iz datoteka čiji tip (format) je potrebno odrediti. Tipično se kao ulazni podaci u algoritme strojnog učenja u ovom problemu ne koriste sami nizovi bajtova, nego histogrami bajtova (učestalost pojavljivanja svakog od 256 bajtova (0 do 255) u danom ulaznom nizu). Za početak istražiti umjetne neuronske mreže, konvolucijske neuronske mreže, algoritam slučajnih šuma, KNN, itd. Odabrati te i još neke algoritme koji su prikladni za klasifikaciju fragmenata datoteka. Obraditi svaki algoritam tako da se objasni barem njegov princip rada, prednosti, ograničenja, tipične primjene i radove u kojima se ta tehnika koristila za klasifikaciju fragmenata datoteka. Fokusirati se ili na širinu područja (upoznati što više algoritama) ili odabrati 5-10 algoritama i za njih pronaći programski jezik ili okolinu u kojoj su već implementirani i istrenirati klasifikator na temelju dostupnog skupa podataka.

Rezultat

Seminar

,

prezentacija

.

Plan rada

1. Pronalaženje iscrpnog popisa algoritama strojnog učenja
2. Početno ispitivanje algoritama i izbor onih koji su prikladni za problem koji se rješava
3. Analiza i opis svakog od odabranih algoritama
4. Identifikacija radova u kojima je taj algoritam korišten za klasifikaciju fragmenata datoteka (pretraga po ključnim riječima koje uključuju ime algoritma, file fragment classification)
5. Fokus na većem broju algoritama (pregled područja u širinu) ili umjesto toga za neki/neke algoritme pronaći programske implementacije i istrenirati klasifikator za podatke.

Vremenski plan rada

Datum	Aktivnost
1. tjedan(18.3.-24.3.)	Upoznavanje teme
2. tjedan(25.3.-31.3.)	Pronalaženje algoritama
3. tjedan(1.4.-7.4.)	Odabir prikladnih algoritama i detaljna analiza
4. tjedan(8.4.-14.4)	Pretraga znanstevnih radova
5. tjedan(15.4.-21.4.)	(19.4.-predaja prve verzije) Pronalazak implementacije određenih algoritama i treniranje

Dnevnik rada

Datum (tjedno)	Aktivnost	Utrošak vremena	Daljnji rad
18.3. - 24.3.	Istraživanje teme. Upoznavanje s neuronskim i konvolucijskim mrrežama, strojno učenje, klasifikacija fragmenata...	7 h	Predaja nacrta.
25.3.-31.3.	Pronalazak algoritama i njihovo proučavanje. 1. Linear Regression 2. Logistic Regression 3. Linear Discriminant Analysis 4. Classification and Regression Trees 5. Naive Bayes 6. kNN	6 h	Detaljnije ispitivanje pronađenih algoritama i ispitivanje koji su prikladni za rješavanje problema.
1.4.-7.4.	Jos neki algoritmi: 7. Learning Vector Quantization 8. Support Vector Machines 9. Bagging and Random Forest 10. Boosting and AdaBoost 11. Principal Component analysis (PCA) 12. Neuronske i konvolucijske mreže https://www.sciencedirect.com/science/article/pii/S1742287613000546#sec1 https://www.sciencedirect.com/science/article/pii/S1742287608000273 https://pdfs.semanticscholar.org/c398/72eae0c61ecf47603aab3f5c1545ee612ac9.pdf http://cs229.stanford.edu/proj2014/Andrew%20Duffy,%20CarveML%20an%20application%20of%	6 h	Pretraga znanstvenih radova i odabir određenih par algoritama za rješavanje problema.
8.4.-14.4.	Detaljno sam proučila neuronske i konvolucijske mreže, kNN i algoritam slučajnih šuma i shvatila njhov princip rada, što primaju, što rade s podacima i što na kraju daju kao izlaz. Odlučila sam te algoritme iskoristit u rješavanju problema te sam našla neke radove u kojima se oni koriste za klasifikaciju fragmenata datoteka. (Nisam ih puno našla tako da mi nije poptpuno jasno kako to točno funkcionira za naš problem...) Započela sam pisanje seminara. https://link.springer.com/content/pdf/10.1007%2F978-3-642-24212-0_5.pdf -za kNN: https://www.researchgate.net/publication/282375639_A_Practical_Video_Fragment_Identification_System -za neuronske: https://www.researchgate.net/publication/303823046_File_Type_Identification_for_Digital_Forensics https://arxiv.org/ftp/arxiv/papers/1002/1002.3174.pdf -za konvolucijske: https://www.researchgate.net/publication/327336441_File_Fragment_Type_Identification_with_Convolutional_Neural_Networks	7h	Istražiti još detaljnije kako ovi algoritmi rade za klasifikaciju fragmenata datoteka. Pronaći još radova. Napisati poglavlja seminara koja sam započela. *Pronaći implementacije za jedan ili više algoritama.
15.4.-21.4.			

Zaključak

Prijedlog za daljnje istraživanje