4. félévi beszámoló

Pataki Bálint Ármin (patbaa@gmail.com)
Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája PhD program
Témavezető: Csabai István

A dolgozat címe: Gépi tanulás a tudományokban

2020.05.29.

Bevezetés:

Főként a digitalizáció hatására napjainkban néhány kivételtől eltekintve minden tudományág nagyságrendekkel több adatot gyűjt, mint akár csak egy-két évtizeddel ezelőtt. Ezzel párhuzamosan a számítási kapacitás is jelentősen növekedett. E két terület fejlődése jelentősen szélesítette a gépi tanulási módszerek alkalmazási lehetőségeit.
Gépi tanulásnak nevezzük azokat a módszereket, amikor egy algoritmus az input adatok alapján tanul meg megoldani valamilyen általunk hasznosnak ítélt feladatot. Leggyakoribb alfaja az ún. felügyelt tanítás, amely során tanító adat-címke párok segítségével próbálja meg egy algoritmus közelíteni az adat→ címke összefüggést a lehető legpontosabban.
Ebbe a szemléletbe feladatok rendkívül széles köre beilleszthető. Amikor orvoshoz megyünk akkor különféle képalkotó módszerekkel készült felvételek (adat) alapján az orvos felállítja a diagnózist (címke). Amikor gravitációs hullámokat szeretnénk detektálni, az szintén felfogható úgy, hogy számos idő-jel adatból szeretnénk meghatározni, hogy az adott időintervallumban a detektorunk észlelt-e gravitációs hullámot.
Ezen feladatok pontosabb, gyorsabb megoldása segíti a tudomány fejlődését, és emellett a mindennapi életünket is. A doktori képzésben eltöltött éveimre azt a célt tűztem ki, hogy tudományos adatokon alkalmazzak gépi tanulási módszereket.

Az előző három félévben elért kutatási eredmények összegzése:

Az előző három félév során számos kutatási projektbe kapcsolódtam be, melyek nagy része még mindig tart.
Kozmológiai paraméterek becslése gravitációsan lencsézett mapek alapján:
A galaxisok fénye mire a távcsöveinkhez ér kissé torzul, rengeteg más hatás mellett, a Föld és a kibocsátó galaxis közötti tér gravitációs mezőjének köszönhetően is. Ezt a gravitációs mezőt a kozmológiai paraméterek segítségével lehet meghatározni. Ily módon különböző kozmológiai paraméterekkel beállított szimulációval le lehet generálni hozzájuk az eredményeket.
A Columbia Egyetem szimulációs adatain tanítottunk konvolúciós neurális hálókat, hogy a végső mapekből minél pontosabban meg tudják határozni, milyen kozmológiai paraméterekkel voltak elkészítve. Hogyha a szimulációk kellően élethűek, akkor ezáltal a jövőben a módszer utat nyithat a valós mérések elemzésével a kozmológiai paraméterek pontosabb meghatározásához.
A kutatásból két publikáció született, melyeken második szerző vagyok.

Antibiotikum rezisztencia meghatározása DNS alapján:
Az antibiotikum rezisztenica növekvő probléma a világban, egyes baktériumok egyre több antibiotikummal szemben ellenállóak. A probléma enyhíthető lenne, hogyha minden antibiotikum kúránál a megfelelő antibiotikumot a megfelelő dózisban alkalmaznánk, hiszen ezen rezisztens baktériumok legtöbbször a nem megfelelő antibiotikum kezelés hatására tudnak elszaporodni. A jelenlegi kísérleti módszerekkel megkapni azt az információt, hogy egy baktérium mely antibiotikumra ellenálló, 1-2 napba

telik. A szekvenálás azonban töretlenül gyorsul, így a jövőben elképzelhető, hogy nagyon rövid idő alatt meg tudunk szekvenálni különböző organizmusokat. A kutatás során *Ecoli.* baktériumok teljes genom szekvenálási adataiból becsültük egy adott antibiotikumra, *ciprofloxacin*, való ellenálló képességét. Három kromoszómális mutáció és egy plazmid gén bizonyult a legnagyobb rezisztencia faktornak.
A kutatásból írt cikk review alatt van a Scientific Reportsban. A cikket túlnyomórészt én szövegeztem, a gépi tanulási és bioinformatikai analízisbe sok munkát fektettem. Első szerző vagyok a cikken.

Kolorektális daganatok felismerése patológiai metszetek alapján:
A SOTE II. sz. Patológiájával együttműködve az ottani orvosok felannotáltak ~2000 digitalizált patológiai metszetet egy közösen kidolgozott módszerrel. Ezen adatokból kiindulva dolgozunk olyan módszereken, amelyek segíthetik a metszeteket kiértékelő orvosokat (előre bejelöli az algoritmus, hogy szerinte hol, milyen elváltozás található). A projektben az elejétől fogva részt veszek, a kezdeti annotációs protokoll kidolgozásán át a neurális hálózatok tanításán és kiértékelésén át a cikkek szövegezéséig.
Két cikket tervezünk publikálni a kutatásból, egy az adatok publikálásával kapcsolatos, a másik pedig az eredmények publikálása lesz. A kéziratok íródnak, de a beküldés idejét még nem látni egyértelműen.

Prosztatarák funkcionális mutációinak vizsgálata:
Spisák Sándorral (Harvard Medical School) kollaborálva olyan nem kódoló DNS szakaszban lévő elváltozásokat kerestünk a humán genomban, amely korrelál a prosztatarák meglétével, és epigenetikailag aktív. Itt szekvenálási eredmények anazílisét, feldolgozását végeztem, a cikken társszerző leszek.
A cikk kézirata előrehaladott állapotban van, várhatóan heteken beküldésre kerül.

Koraszülés valószínűségének becslése génexpressziós adatokból:
Egy gépi tanulási versenyt nyertünk Csabai Istvánnal közösen, mely során a magzat életkorát ill. a koraszülés tényét kellett megbecsülni különböző időpontokban vett vérből származó génexpressziós minták alapján.
A verseny szervezői egy közös cikket írnak az eredményekből, mely kézirata már letisztult, hamarosan beküldésre kerül, melyen társszerző vagyok.

Az aktuális félévben elvégzett kutatások ismertetése:

Az aktuális félév során a fentebb említett orvosi és biológiai témájú kutatási projekteket folytattam, illetve egy újba kezdtünk bele.

Raumatoid arthitis scoring röntgen képek alapján:
Olar Alex végzős MSc hallgatóval közösen belevágtunk egy versenybe, ahol reumatoid artritisz-es páciensek végtagjairól készült röntgenkép alapján kell Keresési eredmények
Internetes találatok Sharp van der Heijde scoret prediktálni, vagyis írületre lebontva meghatározni, hogy milyen szintű és tipusú az ízület károsodása. A verseny a hetekben fog zárulni, jelenleg dobogós helyen állunk. A verseny végeztével egy közös cikket fogunk írni a szervezőkkel, melyen társzerzők leszünk, azonban a cikk beküldésének ideje bizonytalan még.
Emellett felvettük a kapcsolatot magyar reumatológusokkal és radiológusokkal, egy esetleges közös publikáció reményében.

Publikációk:

- Ribli, Dezső, **Bálint Ármin Pataki**, and István Csabai. "An improved cosmological parameter inference scheme motivated by deep learning." Nature Astronomy 3.1 (2019): 93-98.
- Ribli, D., **Pataki, B. Á.**, Zorrilla Matilla, J. M., Hsu, D., Haiman, Z., & Csabai, I "Weak lensing cosmology with convolutional neural networks on noisy data." Monthly Notices of the Royal Astronomical Society 490.2 (2019): 1843-1860.
- Amid C, Pakseresht N, Silvester N, Jayathilaka S, Lund O, Dynovski LD, **Pataki BÁ**, Visontai D, Xavier BB, Alako BTF, Belka A, Cisneros JLB, Cotten M, Haringhuizen GB, Harrison PW, Höper D, Holt S, Hundahl C, Hussein A, Kaas RS, Liu X, Leinonen R, Malhotra-Kumar S, Nieuwenhuijse DF, Rahman N, Dos S Ribeiro C, Skiby JE, Schmitz D, Stéger J, Szalai-Gindl JM, Thomsen MCF, Cacciò SM, Csabai I, Kroneman A, Koopmans M, Aarestrup F, Cochrane G. The COMPARE Data Hubs. Database (Oxford). 2019 Jan 1;2019:baz136. doi: 10.1093/database/baz136. PMID: 31868882; PMCID: PMC6927095.
- Matamoros, S., Hendriksen, R., **Pataki, B.**, Pakseresht, N., Rossello, M., Silvester, N., ... & Schultsz, C. (2019). Accelerating surveillance and research of antimicrobial resistance-an online repository for sharing of antimicrobial susceptibility data associated with whole genome sequences. Microbial Genomics (2020): mgen000342.

Beküldött, elbírálásra váró publikációk:

- SK Sieberts, J Schaff, M Duda, **BÁ Pataki**, M Sun… "Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge." bioRxiv (2020).
- ***Pataki, B. Á.**, Matamoros, S., van der Putten, B. C., Remondini, D., Giampieri, E., Aytan-Aktug, D., ... & Schultsz, C. (2019). Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. bioRxiv, 806760.*

Ezen felül két előrehaladott kézirat van, amelyek várhatóan heteken belül beküldésre kerülnek. Egyik az említett prosztatarákos témában, a másik pedig a terhesség szakaszát génexpresszióból becslő verseny eredményéről.

Tanulmányi tevékenység az aktuális félévben:

Elméleti evolúcióbiológia EA - FIZ/3/005E - még nem kaptam érdemjegyet a beszámoló írásakor

Ezen felül a PhD képzés korábbi féléveiben a szükséges további 7 tárgyat elvégeztem, mindegyiket jeles érdemjeggyel.

Konferenciák a képzés alatt:

- 2019.02.27-03.01. COMPARE General Meeting, Koppenhága, nyelvű előadással vettem részt rajta
- 2019.06.09-12. AICosmo 2019, Ascona, Svájc, nyelvű előadással vettem részt rajta
- 2019.11.04. RECOMB/ISCB Conference on Regulatory & Systems Genomics, New York, előadással vettem részt rajta
- 2019.11.25-26. HEPTECH AIME19 AI & ML, Budapest, itt csak hallgatóság voltam
- 2019.12.09. SOTE Szakmai találkozó a mesterséges intelligenciáról, Budapest, előadással vettem részt rajta

Oktatási tevékenység:

Két szakdolgozat során közreműködtem témavezetőként (egy fizika BSc, és egy TÁTK survey-statisztika MSc).

- 2018/2019/1 Számítógépes alapismeretek (szamalapf18la) - gyakorlat - heti 2 óra
- 2018/2019/1 Adatbányászat és gépi tanulás (dsminingf17vm) - előadás és gyakorlat - heti 3 óra

- 2018/2019/2 A fizika numerikus módszerei I. (fiznum1f18la) - gyakorlat - heti 2 óra
- 2018/2019/2 Deep learning és gépi tanulás a tudományokban (deeplea17em) - előadás - heti 1 óra

- 2019/2020/1 Számítógépes alapismeretek (szamalapf18la) - gyakorlat - heti 2 óra
- 2019/2020/1 Adatbányászat és gépi tanulás (dsminingf17vm) - előadás és gyakorlat - heti 3 óra

- 2019/2020/2 A fizika numerikus módszerei I. (fiznum1f18la) - gyakorlat - heti 2 óra
- 2019/2020/2 Deep learning és gépi tanulás a tudományokban (deeplea17em) - előadás - heti 1 óra

Ezen felül konzultációs/szemináriumi órákon is részt vettem oktatóként, de az nem járt rendszeres, heti előadás/gyakorlat tartásával.

Elismerések:

DREAM Preterm Birth Prediction Challenge, Transcriptomics, nemzetközi gépi tanulási verseny 1. helyezés Csabai Istvánnal csapattársaként.