4. félévi beszámoló

Báskay János (baskayj@student.elte.hu)
Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája PhD program
Témavezető: Pollner Péter

A dolgozat címe: Hálózat-statisztikák adattudományi alkalmazásai

Bevezetés

A 21. század információs forradalma megteremtette a lehetőséget, hogy a természeti és társadalomi jelenségek tanulmányozása során korábban elképzelhetetlennek tűnő adatmennyiséget gyűjthessünk össze. Hamar világossá vált, hogy ezeket manuálisan feldolgozni, majd előrejelzéseket alkotni lehetetlen feladat, erre a problémára nyújtott megoldást a különböző gépi tanulási módszerek megjelenése. Ezek segítségével az adatok közötti korrelációk könnyedén összegyúrhatók egy robosztus modellé, mellyel klasszifikációs vagy regressziós problémák oldhatók meg.

Ezzel párhuzamosan gyors fejlődésen ment keresztül a Hálózattudomány is, melynek keretein belül rendelkezésre áll számos eszköz a komplex rendszerek megértésére és modellezésére.

PhD tanulmányaim során szeretnék a Hálózattudomány eszközeinek felhasználásával és gépi tanulási módszerek interdiszciplináris alkalmazásán keresztül átfogó ismereteket elsajátítani az Adattudomány területén.

Ennek keretében kezdtem el 2020 őszén túlélés analízissel foglalkozni. A túlélés analízis olyan hagyományos statisztikai és gépi tanulási módszereket takar, melyek célja, hogy jóslatokat tegyen arról, hogy egy esemény mikor következik be egy adott mintára, és milyen faktorok játszanak kulcsszerepet az esemény bekövetkezésében. Túlélés analízist az orvostudomány mellett alkalmaznak a mérnöki tudományokban, földtudományokban, szociológiában és pénzügyben is. Regressziós problémáktól az különbözteti meg, hogy nem feltétlenül minden mintán következik be az esemény, azokat a mintákat, melyeken nem fordult elő, cenzorált mintáknak szokás nevezni. Túlélés analízis során használt tipikus statisztikai teszt a LogRank teszt, mely egy χ^2 próba segítségével ellenőrzi, hogy két minta túlélésfüggvénye[1] azonos-e. A túlélés modelleket általában c-index[2] segítségével értékelik ki, amely azt mutatja meg, hogy a modell jóslatainak időrendje megegyezik-e a mintákhoz tartozó ismert időrenddel.

A túlélésanalízissel töltött munkám javát gerincmetasztázisos betegek adatain végeztem, különös tekintettel arra, hogy kiszűrjem a felesleges adatokat (feature selection), és a hasznos adatok a gépi tanulási módszerek számára legjobban feldolgozhatóak legyenek. Emellett besegítettem Mezei Tamás Cox modellen alapuló prognosztikai pontrendszerének publikálásában is, az elkészült modell keresztvalidációjával és egy demo honlap (https://hal.elte.hu/gerincmet/) elkészítésével.

2021 tavasza óta dolgozom a SOTE EMK digitális biomarker programjának keretén belül patológiás metszetek 3D rekonstrukcióján. Ezen munkám során megismerkedtem számos képfeldolgozási, gépi látási (Computer Vision), és képregisztrációs (Image Registration) módszerrel, és részt vettem egy rekonstrukciós módszertan kidolgozásában, melyet jelenleg a SOTE-ELTE közös szabadalomként nyújtott be.

---

[1] $S(x) = 1 - F(x)$, ahol $S(x)$ a túlélésfüggvény és $F(x)$ a kumulatív eloszlásfüggvénye.

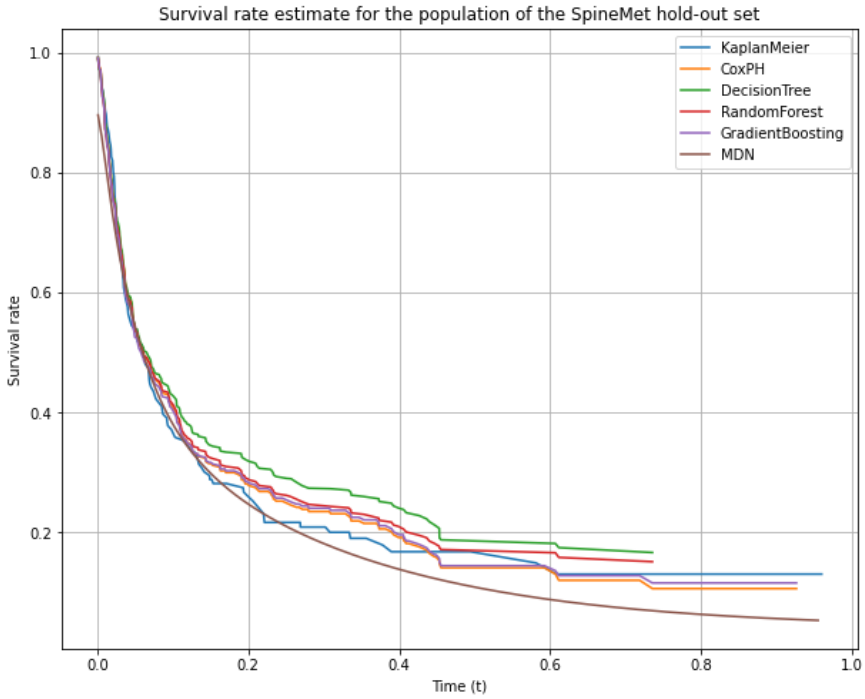
[2] UNO, Hajime, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 2011, 30.10: 1105-1117.

Aktuális félévben elvégzett kutatások ismertetése

Túlélés analízis gerincmetasztázisos betegek adatain gépi tanulással

A félévben célul tűztem ki, hogy a gerincmetasztázisos betegek adatain eddig végzett munkát összefoglaljam és publikációképes állapotba hozzam. Ennek részeként elkészítettem egy túlélés jóslására alkalmas neurális hálót, ami a ”Mixture Density Network” kategóriába sorolható. Ez egy olyan hibrid megoldás, amely tartalmaz egy többrétegű sűrű neurális hálót (MLP – Multi Layer Perceptron), amely egy sűrűségfüggvény keverék modell paramétereit jósolja: A keverési súlyokat, és az egyéni sűrűségfüggvény illeszthető változóit, pl. normális eloszlások esetén az átlagokat és szórásokat. Ilyen megoldást korábban alkalmazott a DWPTE[3] (2021), illetve DeepSurvivalMachines[4] (2021). Az általam javasolt megoldás a kettő keverékének tekinthető, ugyanis használom benne a DWPTE ritka keverék rétegét, amely hivatott kiszelektálni a kevésbé fontos komponenseket a sűrűségfüggvény keverékből, viszont a DWPTE-vel ellentétben nem korlátozom a modellt Weibull eloszlásra, hanem a túlélés analízis irodalmában gyakran használt eloszlások (Exponenciális, Weibull, Gumbel, Logisztikus, LogLogisztikus, Normális, Lognormális és Gamma) mindegyike elérhető[5].

Az általam megalkotott neurális hálót hagyományos, ”sekély” megoldásokkal hasonítottam össze: Cox Proportianal Hazards, Survival Tree, Random Survival Forest és Gradient Boosted Survival Trees. Minden modell paramétereit Bayesi hyperparaméter-optimalizációval hangoltam be, célként pedig 5-Fold keresztvalidációban az Uno-féle c-index maximalizálása volt, feltéve, hogy a modell átmegy a LogRank statisztikai teszten.

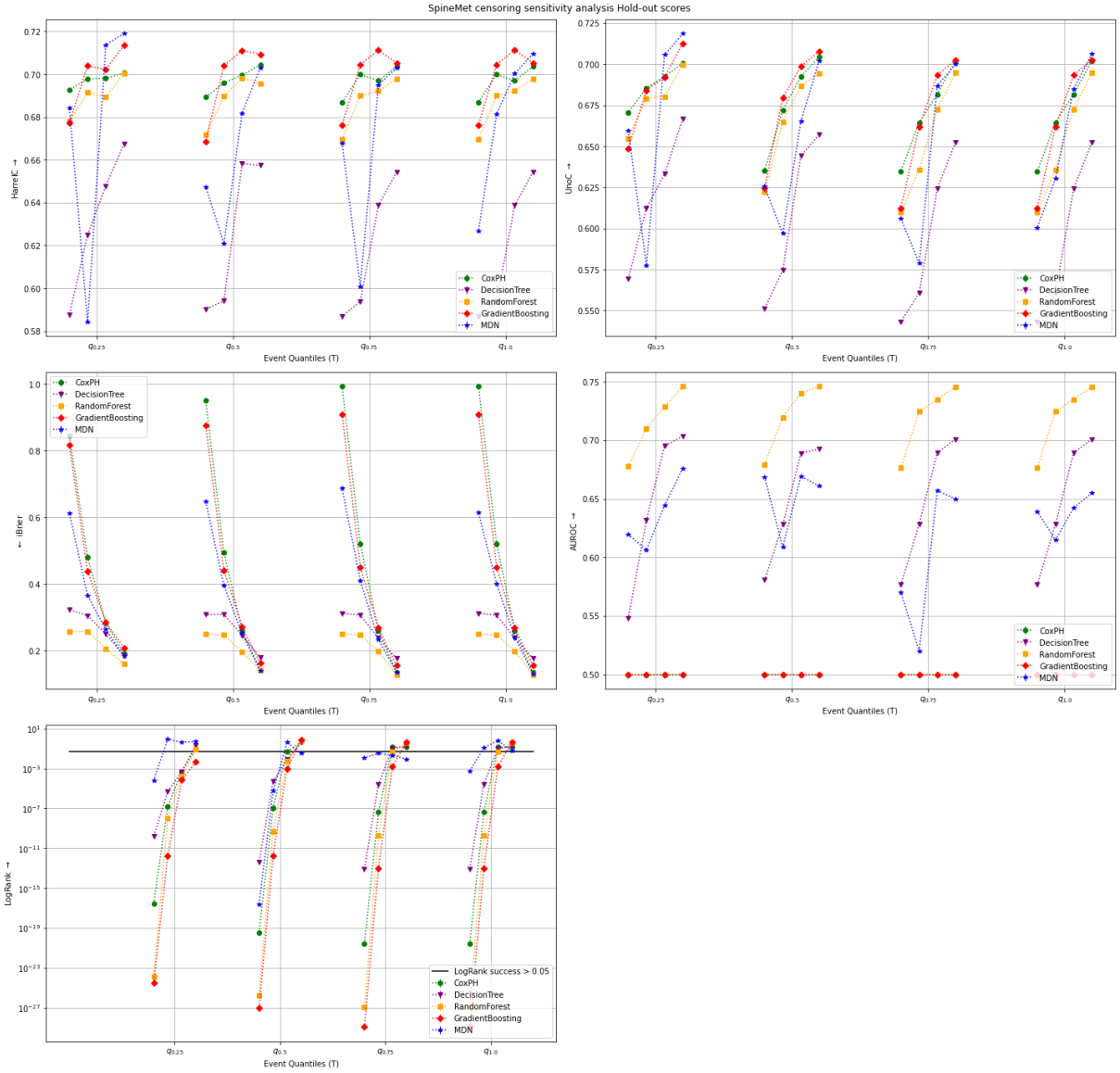


[3] BENNIS, Achraf; MOUYSSET, Sandrine; SERRURIER, Mathieu. DPWTE: A Deep Learning Approach to Survival Analysis Using a Parsimonious Mixture of Weibull Distributions. In: *International Conference on Artificial Neural Networks*. Springer, Cham, 2021. p. 185-196.

[4] NAGPAL, Chirag; LI, Xinyu; DUBRAWSKI, Artur. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 2021, 25.8: 3163-3175.

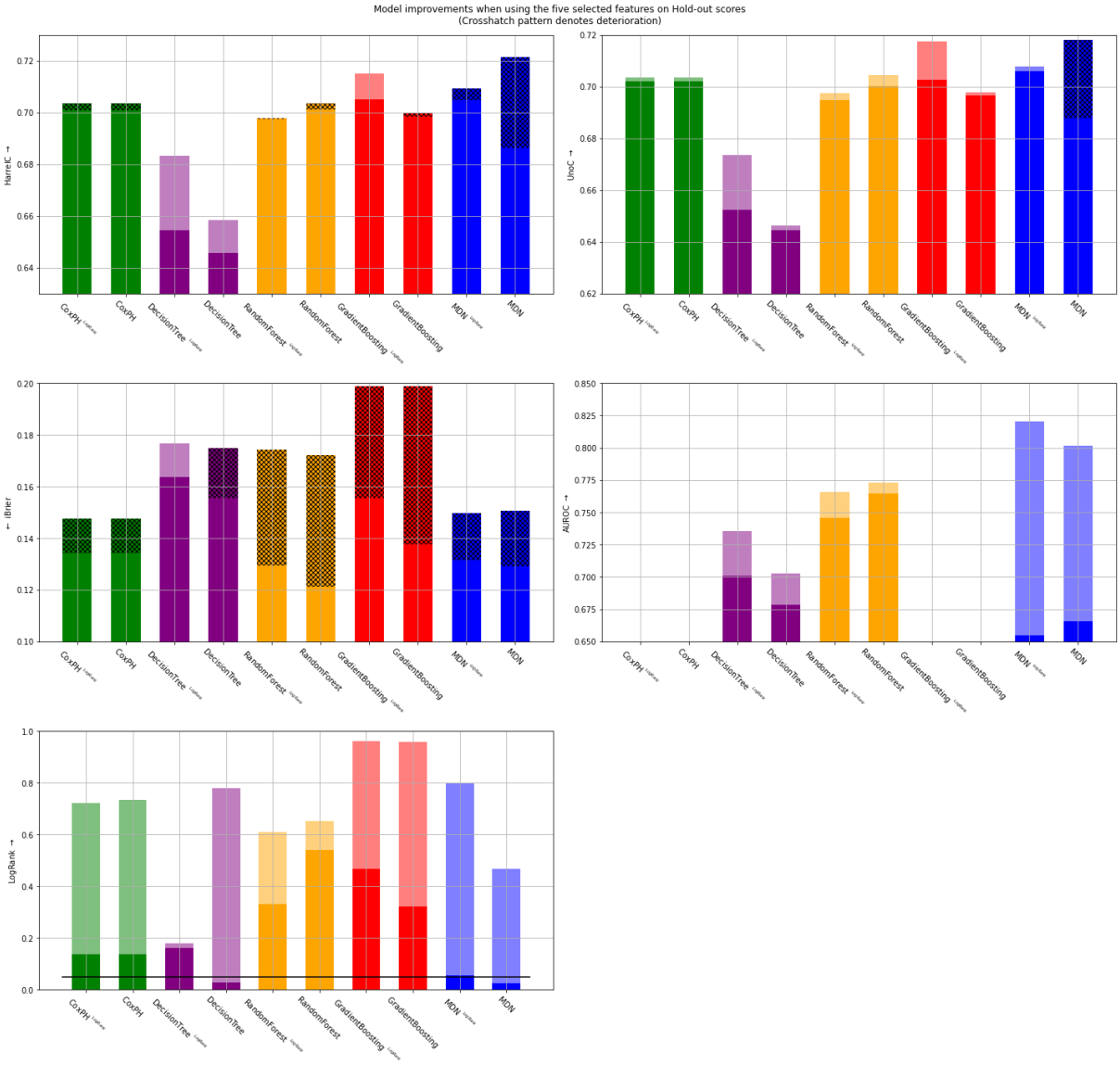
[5] A DeepSurvivalMachines ezzel szemben jelenleg mindössze Normális, Weibull és Logisztikus eloszlások keverékét tudja illeszteni.

A kapott modelleket ”censoring sensitivity analysis”-nek vetettem alá, melynek lényege, hogyha romlik a tanító halmaz minősége (egyre kevesebb mintán következik be az esemény), akkor a különböző modellek mennyire képesek jó előrejelzéseket tenni. Ezek alapján arra jutottam, hogy noha a keresztvalidáció során a fa alapú ensamble modellek produkálják a legjobb c-indexet, a LogRank teszten már kevésbé mennek át az adatok minőségének romlásával, ami azt is jelenti, hogy a jósolt túlélések már nem tartoznak ugyanabba az eloszlásba, mint a valódiak.



További szembetűnő eredmény, hogyha a teljes adatszett helyett mindössze azt az 5 tulajdonságot használom a modellek tanítására, melyet az Ideggyógyászati Szemlében[6] megjelent pontrendszer esetén használtunk, A LogRank teszten az összes modell teljesítménye jelentősen javul, de a c-indexeken is (kisebb) növekedés tapasztalható, bár ennek az az ára, hogy az integrált Brier pont egy modellt kivéve mindenhol romlik. Ez az eredmény indokolja, hogy a jövőben érdemes lesz különösebb gondot fordítani egy túlélés analízissel kompatibilis feature szelekciós módszertan kidolgozására is, ami remélhetőleg még pontosabb és flexibilisebb modellekhez vezet majd.

[6] MEZEI, Tamás, et al. New, innovative prognosis calculator for patients with metastatic spinal tumors. *Ideggyogyaszati Szemle*, 2022, 75.3-04: 117-127.



Ezzel a kutatás eljutott egy olyan pontra, hogy az eredményeket érdemes egy publikációban összefoglalni. Ehhez az illusztrációk és irodalomjegyzék már elkészült, a kézirat megírása folyamatban van.

3D képrekonstrukció patológiás metszeteken gépi tanulás segítségével

A szemeszter folyamán számos megbeszélés volt a SOTE Innovációs Központtal a módszertan tervezett szabadalmának benyújtása kapcsán, kértünk előzetes oltalmazhatósági véleményt is, mely összességében pozitív volt. A szabadalom májusban beadásra került, így júniusban a cikk megjelentetése is megkezdődhet.

Publikációk

- **Mezei, T., Báskay, J., Pollner, P., Horváth, A., Nagy, Z., Czigléczki, G., & Banczerowski, P.** (2022). New, innovative prognosis calculator for patients with metastatic spinal tumors. *Ideggyogyaszati Szemle*, 75(3-04), 117-127.
- **Báskay, J., Kivovics, M.,Pénzes, D., Kontsek, E., Pesti, A., Szócska, M., Németh, O., Pollner, P.** (----) Reconstructing 3D histological structures using machine learning (AI) algorithms.: Várhatóan a *Scientific Reports*ban jelentetjük meg, ha a hozzá

tartozó szabadalom a megfelelő fázisba került. A kézirat 2021. novembere óta készen van.

- Előkészülőben van továbbá egy cikk, ami összefoglalja a gerincmetasztázisos betegek adatain gépi tanulással végzett túlélésanalízist. Ábrák, irodalomjegyzék már vannak, kézirat készülőben. Szintén a *Scientific Reports*ba szánjuk.

Tanulmányi tevékenység az aktuális félévben

- Elméleti evolúcióbiológia (FIZ/3/0005E)
- A gépi tanulás új eredményei szeminárium (FIZ/3/092)

Oktatási tevékenyég az aktuális félévben

A mesterszakos Fizika hallgatók és doktoranduszok számára meghirdetett *Haladó statisztika és modellezés* tárgy keretén belül túlélés analízisről tartottam előadást és gyakorlatot.