

Bevezetés

A felügyelt gépi tanulás során célunk, hogy egy jól definiált eredményváltozóra minél nagyobb pontosságú becslést adjunk bizonyos magyarázóváltozók értékének ismeretében. Napjainkban a feladat számtalan összetett algoritmus segítségével megoldható. Pl. mélytanuló neurális hálózatok, véletlen erdők stb. Azonban egyre több szerző, pl. Molnar (2020) hívja fel a figyelmet arra, hogy a legpontosabb becslést szolgáltató modellekben a használt magyarázóváltozók hatásai az eredményváltozóra nehezen, vagy egyáltalán nem visszafejthetők. Viszont, bizonyos gyakorlati szituációkban a gépi tanulás legfontosabb eredménye nem feltétlenül a minél pontosabb becslés elkészítése, hanem az egyes magyarázóváltozók hatásának megállapítása. Például, egy banknak egyértelműen meg kell indokolnia, hogy mi alapján utasít el egy hitelkérelmet. Ilyen esetekben nem előre jelző, hanem magyarázó modellek építése az elemző célja.

Az általánosított additív modellek (az angol Generalized Additive Model kifejezésből GAM-nak rövidítve) átmenetet képeznek a modell predikciós pontossága és az értelmezhetőség között:

$$h(E(Y)) = \varepsilon + \sum_{j=1}^p f_j(X_j)$$

A magyarázóváltozók hatása nem-lineáris, de additív módon elkülönülő, így a f_j függvény segítségével könnyen értelmezhetők.

Kutatási kérdés

Napjaink „big data” környezetében, amikor egy adott becslési feladathoz rengeteg potenciális magyarázóváltozó könnyen az elemző rendelkezésére áll, még egy egyszerű lineáris regressziós modell alkalmazása esetén is problémás lehet a magyarázóváltozók hatásainak megállapítása. Molnar (2020) egyik javaslata a probléma áthidalására, a változószelekció.

A kutatás során egy saját fejlesztésű hibrid genetikus – harmónia kereső metaheurisztikus algoritmus (továbbiakban Hibrid Algoritmus, HA) alkalmazását vizsgáljuk meg GAM-ok változószelekciós feladatára. A feladatba felveszünk egy olyan korlátot, ami a magyarázóváltozók közti redundancia elkerülését biztosítja. A korlátozott változószelekciós feladatot két valós adatbázison oldjuk meg.

Az eredményeink alapján a metaheurisztikus megoldásunk hatékonyabban tudja kezelni a korlátot, mint a szakirodalom által javasolt két korszerű algoritmus. Az algoritmus hatékonyságát jelentősen befolyásolja a véletlen és szabályszerű szelekciós operátorok közti optimális kombináció megtalálása.

Nagyobb adatbázison az algoritmus futásideje jelentős, ám párhuzamosítással egy véletlen erdő alapú megoldással összemérhető.

Módszertan

Hall (1999) szerint az optimális magyarázóváltozók

1. összefüggnek a célváltozóval és
2. nem redundánsak.

GAM-ok esetében is alkalmazható változószelekciós algoritmusok közül csak kettő helyez hangsúlyt a 2. pontra is:

- Minimum Redundancy Maximum Relevance (mRMR) (Song et al., 2012)
- HSIC – Lasso (Climente-González et al., 2019)
- Csak a magyarázóváltozók páronkénti függetlenségére kontrollálnak!

Saját fejlesztésű HA-ban egy magyarázóváltozó nem lehet a többi használt változó többváltozós függvénye sem.

A három algoritmus teljesítményét két valós adatbázison hasonlítjuk össze.

Betongerendák nyomószilárdságának becslése a gerendák hét összetevője és a koruk függvényében. 1030 db betongerenda 9 ismerv. Kis feladat, ismert optimum, cél a paraméterek finomhangolása.

Tajvani bank ügyfeleire becsljük meg, hogy az ügyfél a lekérdezés időpontjától számított 1 hónapos időtartamon belül csődöt fog-e jelenteni hitelkártya adósságára. Az adatbázis 30 000 rekordot tartalmaz, 26 lehetséges magyarázóváltozóval. Nagy méretű, életszerű feladat.

Mindkét adatbázist 70%-30% arányban osztottuk fel tanító- és tesztmintákra. A vizsgált algoritmusok implementálása és futtatása az R 3.5.3 verziójában történt. Kivételt képez a HSIC-Lasso algoritmus, aminek jelenleg csak Python implementációja elérhető. Ez utóbbi implementációt a Repl.it online fejlesztőkörnyezetben futtattuk.

Az algoritmusokat a véletlen elemeik miatt 30-szor futtatjuk le az adatbázison, és a legjobb megoldásokat hasonlítjuk össze.

Eredmények - Betongerendák

Algoritmus	Kiválasztott változók	Redundáns változók	R ² (Tesztminta)
mRMR	cement, pernye, víz folyósítóanyag, finom aggregátum, kor, beton	cement, pernye, víz folyósítóanyag, beton	82.123%
HSIC-Lasso	cement, salak, víz, folyósítóanyag, kor, beton	víz, folyósítóanyag	86.382%
Hibrid metaheurisztika	cement, salak, víz, kor	-	84.479%

A redundanciák oka, hogy az építőiparban alkalmazott gerendákban a folyósítóanyag mennyisége a víz/cement arány függvénye.

Így ez a változó redundáns, ha a másik kettő is a modellben szerepel.

Eredmények – Hitelkártya csőd

- PAY_X változók:
 - ügyfél hátralékos hónapjainak száma lekérdezés előtt X hónappal
- PAY_AMT_X változók:
 - lekérdezés előtt X hónappal fizetett törlesztőrészlet összege

Benchmark modellek alkalmazása: döntési fa és véletlen erdő (Recursive Feature Elimination (RFE) változószelekcióval).

Algoritmus	Kiválasztott változók	Redundáns változók	AUC (Tesztminta)	Átlagos futásidő (perc)	Futásidők szórása (perc)
mRMR	AGE, PAY_0+2+5, PAY_AMT_1+6	PAY_2	0,7441	1,629	0,035
HSIC-Lasso	AGE, PAY_0+2+3+4	PAY_0+2+3	0,7409	0,446	0,048
Hibrid algoritmus	LIMIT_BAL, PAY_0, PAY_AMT3	-	0,7488	147,257	15,079
Döntési fa	PAY_0	-	0,6413	0,0127	0,0004
Véletlen erdő (RFE szelekcióval)	mindegyik	LIMIT_BAL, PAY_0+2+3+4+5+6, BILL_AMT1+2+3+4+5+6, PAY_AMT1+2+3+4+5+6	0,7562	132,543	0,758

Redundancia oka, hogy a fizetési státusz a lekérdezés hónapjában jól közelíthető a korábbi hónapok fizetési státuszának többváltozós függvényével.

Hibrid algoritmus teljesítménye a tesztmintán nem látványosan marad el a véletlen erdőtől, és ezt a teljesítményt mindössze 3 változóval éri el 26 helyett.

Futásidőben is összemérhető a HA az RFE változószelekcióval kombinált véletlen erdővel a **párhuzamosítás után: 9,8 óráról 2,5 órára csökkenthető a várható futásidő.**

Következtetések

Mindkét vizsgált adatbázison a HA végső modellje a legnagyobb becslési pontossággal rendelkező modell, amely változói nem redundánsak.

Amennyiben felügyelt tanulás során a cél egy takarékos, magyarázó jellegű modell építése, és a futásidő sem komoly korlát, akkor a HA alkalmazása javasolt az egyéb vizsgált változószelekciós algoritmusokkal szemben GAM keretben.

Az algoritmus hatékonyabb, ha a véletlenszerű szelekciós operátorai dominálnak, de a szabályszerű elemek teljes elhagyása nem kifizetődő.

Az algoritmus várható futásideje párhuzamosítással látványosan javítható.

Későbbi munkáinkban tervezzük a HA eredményeinek összevetését más változószelekciós algoritmusokkal. Vizsgálni kívánjuk a HA kompatibilitását egyéb (nem GAM) modellkeretekkel is.

Irodalomjegyzék

Climente-González, H., Azencott, C. A., Kaski, S., & Yamada, M. (2019). Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14), i427-i435.
Hall, M. A. (1999). Correlation-based feature selection for machine learning. *Doktori értekezés*. University of Waikato.
Molnar, C. (2020). *Interpretable machine learning*. Leanpub.
Song, L., Smola, A., Gretton, A., Bedo, J., & Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 1393-1434.