Alkalmazott Matematikai Lapok **41** (2024), 1–15.
DOI: xxx

PARAMÉTERES AKTIVÁCIÓS FÜGGVÉNY ÉS JELENTŐSÉGE A NEURÁLIS HÁLÓK ÉRTELMEZHETŐSÉGÉBEN

PUSZTAHÁZI LUCA SÁRA, EIGNER GYÖRGY, CSISZÁR ORSOLYA

A mesterséges intelligencia, különösen a mélytanulási modellek forradalmasítják az üzleti és technológiai világot. Napjaink egyik legnagyobb kihívása a mélytanulásban az értelmezhetőség problémájának megoldása. A modellek átláthatóságának, teljesítményének és biztonságának javítására (XAI: eXplainable Artificial Intelligence) egyre égetőbb szükség van. A neurális hálózatok folytonos logikával és többkritériumú döntéshozatali eszközökkel való kombinálása hozzájárulhat a jobb értelmezhetőséghez, átláthatósághoz és biztonsághoz az orvosi, mérnöki és üzleti alkalmazásokban. Ez a megközelítés más fejlődő módszerekkel együtt a neuroszimbolikus hibrid mesterséges intelligenciához, a mesterséges intelligencia kutatásának újszerű területéhez tartozik, amely a hagyományos szabályalapú megközelítéseket modern mélytanulási technikákkal kombinálja. A neuroszimbolikus modellekről bebizonyosodott, hogy a hagyományos modellekhez képest lényegesen kevesebb adattal is nagy pontosságot érnek el. A neurális hálózatok és a szimbolikus rendszerek kiegészíthetik egymás erősségeit és gyengeségeit, így pontos, mintavételi szempontból hatékony és értelmezhető rendszerek jöhetnek létre. Ez javíthatja a döntéshozatalt, bizalmat építhet a gépi tanulási modellek iránt, és hatékonyabb és eredményesebb folyamatokhoz vezethet a különböző iparágakban. Ebben az összefoglaló cikkben kutatócsoportunk e területen elért legújabb eredményeit ismertetjük.

1. Bevezetés

Az értelmezhetőség igénye a gépi tanulásban egyre nagyobb szerepet kap. A hagyományos neurális hálók gyakran nagyszámú paramétert tartalmaznak, így meglehetősen nehéz megfogalmazni, hogy miért hoz egy neurális háló bizonyos döntéseket. Számos, valós életben található alkalmazási területen azonban elengedhetetlen, hogy az eredményül kapott döntés értelmezhető legyen. Néhány intuitív példa:

1. Az orvosok nem szeretnék, hogy neurális hálók is részt vegyenek a kezelések felírásában, amíg nem értik teljesen pontosan, hogy mely tünetek kombinációja indokolja az adott kezelést.

2. A biztosítók modelljeit ellenőrizni kell annak érdekében, hogy egyetlen kisebbséget se érjen hátrányos megkülönböztetés.

3. A robotfejlesztőknek érteniük kell, hogy adott környezeti feltételek mellett miért hoznak a robotok bizonyos, akár látszólag ellentmondásos döntéseket.

4. Amikor egy önvezető jármű balesetet szenved, mind jogi, mind mérnöki szempontból fontos ismerni a jármű döntéseinek indokait.

Mindegyik esetben értelmezhetőbbé válna a modell, ha a neurális hálózatot logikai kifejezésként tudnánk leírni.

A folytonos logika bevonása ígéretes út az értelmezhető modellek irányába, hiszen ahhoz, hogy a számítások természetes nyelven megfogalmazhatók legyenek, jó szolgálatot tesznek a pontatlan, nem teljesen precíz kifejezések. A szakértők által megfogalmazott szabályok digitális nyelvre történő „lefordításához" és a természetes nyelv, valamint az emberi gondolkodás jobb modellezése érdekében dolgozta ki Lotfi Zadeh a fuzzy logika alapjait. Zadeh a korai 1960-as években megfigyelt egy különös jelenséget: a szakértők általi vezérlés jobb eredményekhez vezet, mint az optimális automatikus vezérlés. Jobban meggondolva érthető a jelenség, mivel a szakértők olyan további tudást is használnak, amit nem lehetett az automatikus vezérlőkbe beprogramozni. Ezt a további tudást pontatlan nyelvi elemekkel lehet megfogalmazni, mint például, „ha kicsit csökken a nyomás, akkor kicsit több anyagot kell beleengedni az edénybe". Itt a „kicsit" szónak nincs pontos jelentése. A Zadeh által kidolgozott fuzzy logika segítségével ez a pontatlan tudás leírható precíz matematikai eszközökkel.

A neurális hálók és a folytonos logika összekapcsolása egyre nagyobb érdeklődésnek örvend. Az adaptív folytonos logikai rendszereket évtizedek óta tanulmányozzák [1, 2, 3], a neurális folytonos modellezés pedig továbbra is aktív kutatási téma [4, 5, 6]. Ezen módszerek ötvözésének a fő előnye az, hogy egy olyan modellt eredményez, amely rendelkezik a hagyományos, „fekete doboz" típusú neurális hálók rugalmasságával és pontosságával, valamint a folytonos logikai rendszerek értelmezhetőségével és átláthatóságával. A mély neurális hálózatokban az adatfeldolgozás általában lineáris transzformációk és az $f(x) = \max(0, x)$ rektifikált lineáris egység (ReLU) váltakozó végrehajtásával történik. Érdekes módon mindkét művelet, $f_{\&}(a, b) = \max(a + b - 1, 0)$ és $f_{\vee}(a, b) = \min(a + b, 1)$, könnyen ábrázolható ilyen neurális hálózati transzformációk kompozíciójaként. Ez a tény segít értelmezni a mély neurális hálózatokban lévő transzformációkat „és"- és „vagy"-műveletek formájában – így lehetőséget adva arra, hogy a mély neurális hálózatok empirikus sikerét kiegészítsük az eredmények természetes nyelvi interpretációjával [7, 9].

Az emberi viselkedés tanulmányozása fontos összetevője több tudományterületnek, például a számítástudománynak, mesterséges intelligenciának, neurális hálózatoknak, kognitív tudománynak, filozófiának és pszichológiának. Gyakran feltételezzük, hogy a viselkedést az észlelés, a tudás és a mentális folyamatok befolyásol-

ják. Az emberi viselkedés modellezéséhez olyan számítógépes-logikai rendszerekre van szükség, amelyek képesek magas szintű gondolkodást kezelni. Ilyen rendszerek közé tartoznak a klasszikus logika, a nemmonoton logika, a modális és temporális logika. Az emberi viselkedés modellezésében a gépi tanulás alapú modellek közül elsősorban a kapcsolati modellek, mint a feedforward és rekurrens hálózatok, szimmetrikus és mély hálózatok, valamint az önszervező hálózatok játszanak nagy szerepet. Azok a modellek, amelyek foglalkoznak a kognitív folyamatok valószínűségi jellegével, például a Bayes hálók, a Markov döntési folyamatok és a valószínűségi logikai modellek, szintén fontosak a viselkedés modellezésében a bizonytalanság kezelésére [10].

Jelen cikkben röviden bemutatjuk egy hibrid neuroszimbolikus modell matematikai hátterét és kutatócsoportunk e területen elért legújabb eredményeit.

2. Neurális hálók és a folytonos logika

A gépi tanulásban a számítási idő minimalizálása nagy adathalmaz esetén természetesen elengedhetetlen. Ideális esetben azonban a számításoknak értelmezhetőnek is kell lenniük, mivel akár az időjárás, akár a banki kölcsön esetén meg kell magyarázni a kérdező számára, hogy miért az adott eredményt adta a matematikai modell.

A gyors számításokra vonatkozó igény természetes úton vezet a neurális hálók használatához, mivel azok jellegüknél fogva párhuzamos műveleteket tartalmaznak, amelyek nagyban hozzájárulnak a számolás sebességének növeléséhez. A függvényválasztás tekintetében a lineáris függvények a leggyorsabban számolhatók, a nemlineáris függvények közül pedig az egyváltozós $y = s(z)$ alakúak. Ezáltal eljutunk a neurális hálók réteges felépítéséhez, melyben minden réteg tartalmaz egy lineáris transzformációt és egy, a lehető leggyorsabban kiszámolható nemlineáris transzformációt, és a rétegek egymás után számolandók. A nemlineáris függvényt ebben az esetben aktivációs függvénynek hívjuk. Az aktivációs függvények közül hagyományosan a legelterjedtebb a szigmoid ($s(z) = 1/(1 + \exp(-z))$) függvény. Az utóbbi időben azonban egyre sikeresebb a ReLU ($s(z) = \max(0, z)$) függvény, amely a negatív bemeneteket nullára korrigálja, míg a pozitív bemeneteket változatlanul hagyja.

Az értelmezhetőség igénye természetes úton vezethet érvelés, valamilyen típusú logika irányába. A folytonos logika a hagyományos, kétértékű logikával ellentétben, ahol minden kijelentés vagy igaz, vagy hamis (1 és 0 számokkal reprezentálva), bizonyossági fokokkal dolgozik, amelyek köztes értékeket vesznek fel a [0,1] intervallumban. Ezzel a megközelítéssel a klasszikus logikánál jobban modellezi az emberi gondolkodást.

A logikai neuronokon alapuló neurális háló elterjedt kutatási téma a széleskörű alkalmazhatóság és értelmezhetőség miatt. Ezek a modellek kiválóan alkalmazha-

tók például mintázatok klasszifikációjára, idősorok előrejelzésére, árverési csalások kiszűrésére, illetve nemlineáris folyamatok elemzésére. A logikai neuronok nemlineáris leképezések $[0,1]^m \to [0,1]$ között, ahol az „és” neuronok leggyakrabban t-normákat, míg a „vagy” neuronok t-konormákat (s-normákat) reprezentálnak.

[16]-ban a szerzők ezért azt elemzik, hogy a folytonos logika mely „és” és „vagy” operátorai reprezentálhatók a leggyorsabb, egy rétegű neurális hálóval, illetve hogy mely aktivációs függvény használható az adott reprezentáció esetén. Megmutatható [16], hogy az egyetlen olyan „és” és „vagy” operátorok, amelyeket egy egy rétegű neurális hálóval ki lehet számítani, azok az

$$f_{\&}(a,b) = \max(a+b-1, 0), \tag{1}$$

valamint

$$f_{\vee}(a,b) = \min(a+b, 1) \tag{2}$$

függvények, a ReLU függvény pedig az az aktivációs függvény, amellyel ezek a leggyorsabban kiszámíthatók. Ez a megfigyelés a függvény növekvő népszerűségét is magyarázza.

Egy további érdekes tulajdonság is alátámasztja a fenti operátorok használatát. A folytonos logikában az ellentmondás törvénye (amely szerint x és *nem* x mindig hamis) és a kizárt harmadik törvénye (x vagy *nem* x mindig igaz) nem mindig teljesülnek. Azokban a rendszerekben, amelyekben az „és”, „vagy” és „nem” műveletek hármasa kielégíti a fenti két törvényt, az „és” és „vagy” operátorok éppen a fenti ((1), (2)) függvényekkel izomorfak [7, 8].

3. Uninormák és a paraméteres aktivációs függvény

Az emberi gondolkodásban a logikai operátorokon kívül a kevert operátorok is fontos szerepet játszanak, ahol a magas bemeneti értékek kompenzációt nyújthatnak az alacsonyabb bemeneti értékeknek. Emiatt Yager és Rybalov [17] bevezették az uninorma fogalmát, amely általánosítja a t-normát és a t-konormát (s-normát). Azóta egyre több kutatás szól az uninormákról mind elméleti, mind gyakorlati szempontból, melynek során hasznosnak bizonyultak több területen is, mint például szakértői rendszerek és folytonos logikai integrálok esetén.

A folytonos logikai „és” és „vagy” műveleteket modellező neuronok mellett ezért új fajta kapcsolók, mint például az uninormák és nullnormák használata is egyre jobban terjed. Az uninormák alacsony bemeneti értékek esetén konjunkcióként, magas bemeneti értékek esetén diszjunkcióként viselkednek, míg a nullnormák éppen fordítva. Vegyes bemeneti értékek esetén mindkét típus átlagoló függvényként viselkedik, így fontos szerepet játszanak a többváltozós döntési modellekben.

3.1. Definíció. (Uninorma, Yager és Rybalov, 1996) *Uninormának* hívjuk azt az $U : [0,1] \times [0,1] \to [0,1]$ leképezést, amely kommutatív, asszociatív, nemcsökkenő és létezik $e \in [0,1]$ neutrális eleme, melyre $U(e,x) = x$ minden $x \in [0,1]$ esetén teljesül.

3.2. Definíció. (Aggregatív operátor, Dombi, 1982) *Aggregatív operátornak* nevezzük azt az $a : [0,1] \times [0,1] \to [0,1]$ leképezést, amelyre a következők teljesülnek:

- folytonos a $[0,1]^2 \backslash \{(0,1),(1,0)\}$ halmazon;
- $a(x,y) < a(x,y')$, ha $y < y', x \neq 0, x \neq 1$;
  $a(x,y) < a(x',y)$, ha $x < x', y \neq 0, y \neq 1$;
- $a(0,0) = 0$ és $a(1,1) = 1$ (szélsőérték feltételek);
- létezik egy erős tagadás η, melyre $a(x,y) = \eta(a(\eta(x),\eta(y)))$, ha $\{x,y\} \neq \{0,1\}$ (self-De Morgan identity);
- $a(1,0) = a(0,1) = 0$ vagy $a(1,0) = a(0,1) = 1$.

A fő különbség az uninormák és az aggregatív operátorok definíciója között az, hogy az aggregatív operátorok önduális feltétele nem jelenik meg az uninormák definíciójában, míg az uninormák neutrális elem tulajdonsága nincs benne az aggregatív operátorok definíciójában. [7]-ben a szerzők vizsgáltak egy általános paraméteres keretrendszert a nilpotens konjunktív, diszjunktív, aggregatív és negáló operátorokhoz és megmutatták, hogyan használható az aggregatív operátor preferenciák modellezésére. [15]-ben a szerzők ezt felhasználva építenek egy modellt, ahol az aktivációs függvény egy általános operátorként szerepel, amely tartalmazza a bemeneti értékek közötti kompenzáció mértékét jelölő tanítható paramétert.

3.3. Definíció. (Vágófüggvény) Definiáljuk a *vágófüggvényt* a következő formulával:

$$[x] = \begin{cases} 0, & ha \quad x < 0, \\ x, & ha \quad 0 \leq x \leq 1, \\ 1, & ha \quad 1 < x. \end{cases}$$

3.4. Definíció. (Súlyozott általános operátor) Legyenek $\mathbf{w} = (w_1, \ldots, w_n)$ és $w_i > 0$ valós paraméterek, $f : [0,1] \to [0,1]$ monoton növekvő bijekció és $\nu \in [0,1]$. A *súlyozott általános operátor* a következő formulával definiálható:

$$a_{\nu,\mathbf{w}}(\mathbf{x}) := f^{-1}\left[\sum_{i=1}^{n} w_i \left(f\left(x_i\right) - f(\nu)\right) + f(\nu)\right].$$

A súlyozott általános operátort az implementált modellben két változóval, konstans együtthatókkal és az identitás függvénnyel használjuk, $\mathbf{x} = (x, y)$; $\mathbf{w} = (w_1, w_2) = (1, 1)$; $f(x) = x$; azaz

$$a_\nu(\mathbf{x}) = [x + y - \nu].$$

A fenti operátor definíciójában szerepel a vágófüggvény, mely nem mindenhol differenciálható. Viszont a neurális háló gradiens alapú optimalizáló módszerek segítségével tanítható, melyhez szükség van az aktivációs függvény deriváltjára. Ennek érdekében bevezetjük a vágófüggvény jelenleg ismert legjobb közelítését, a squashing function-t, melyet az alábbi képlettel definiálhatunk [11, 12, 13]:

$$S_{a,\lambda}^{(\beta)}(x) = \frac{1}{\lambda\beta} \ln \frac{1 + e^{\beta(x-(a-\lambda/2))}}{1 + e^{\beta(x-(a+\lambda/2))}}.$$

Az implementációban a következő paraméterekkel használjuk: $\lambda = 1$, $a = \frac{1}{2}$, $\beta = 80$, azaz

$$S(x) = \frac{1}{80} \ln \frac{1 + e^{80x}}{1 + e^{80(x-1)}}.$$

A paraméteres aktivációs függvény így a következő: $S(x + y - \nu)$, melyben ν a tanítható paraméter.

1. táblázat. A súlyozott általános operátor három speciális esete

Típus	ν	$\mathbf{a}_\nu(\mathbf{x}, \mathbf{y})$
Diszjunkció	0	$[x + y]$
Konjunkció	1	$[x + y - 1]$
Átlagoló operátor	0.5	$[x + y - 0.5]$

A paraméteres operátor így már differenciálható és ezáltal alkalmazható aktivációs függvényként. A neurális háló tanítása során a paraméteren keresztül az optimális logikai operátort is tanulja a rendszer. A paraméter értéke jelöli a bemenetek közötti kompenzáció mértékét, mely szerint lehet konjunktív, diszjunktív vagy átlagoló operátor, amint az 1. táblázatban is látható.

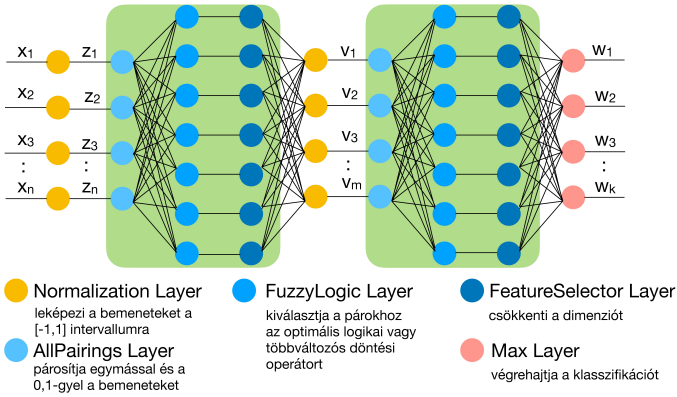
Gashler és társai [14] egy innovatív, paraméterezhető, differenciálható aktivációs függvényt használtak a mélytanulási architektúrában fuzzy logikai kifejezések tanulására. Ez az aktivációs függvény lehetővé teszi a neurális hálózat számára, hogy a bemeneti változók közötti kapcsolatokat megtalálja.

Kutatócsoportunk a Csiszár és Dombi által [7, 9]-ben megalapozott elméleti alapok alapján a Gashler és társai [14]-ben bemutatott modellt fejlesztette tovább [15]. Egy olyan hibrid neuroszimbolikus modellt alkalmaztunk, amelyben a logikai operátorok és egy uninorma tulajdonságú kompenzáló operátor egyetlen

paraméteres függvény segítségével írható le. A tanulható paraméter egy paraméteres aktivációs függvénybe kerül. Így a neurális háló tanítása során az aktivációs függvényen keresztül a logikai operátorok paraméterei is finomodnak mindaddig, amíg a rendszer megtanulja az optimális műveletet a bemeneti értékek között.

4. A modell alkalmazása klasszifikációs problémákon

A bemutatott mély tanulási modell célja, hogy az emberek által érthető kapcsolatokat találjon a bemeneti tulajdonságok között. A fenti megközelítés egy paraméteres, differenciálható aktivációs függvényt használ, amely a nilpotens logikai rendszerek súlyozott általános operátorán alapul.



1. ábra. A neurális háló felépítése

A neurális háló implementációja [14, 15] 9 rétegből áll (1. ábra), a bemeneti értékeket először normalizálja a modell, ami azt jelenti, hogy leképezi a $[-1, 1]$ intervallumra. A következő réteg létrehozza a normalizált értékekből és az Igaz, Hamis értékekből az összes lehetséges párosítást. A harmadik rétegben a fenti aktivációs függvény segítségével a modell megtanulja az optimális folytonos logikai operátort mindegyik párosításhoz. A negyedik réteg csökkenti a dimenziót azáltal, hogy szelektál a bemeneti értékek között. A 2-4. rétegek együtt a neurális háló logikai része, ennek többszörös alkalmazásával mélyíthető a modell, jelen esetben két logikai részt tartalmaz. Két logikai rész között szükséges az értékek $[-1, 1]$-re transzformálása a tanh függvény segítségével. Az utolsó, kilencedik réteg a klasszifikációért felel a max függvényt használva.

A modell (Fuzzy NN) validálása a UCI Machine Learning Repository-ból vett klasszifikációs problémákon történt. Referenciaként egy hagyományos mély ne-

2. táblázat. Hibás besorolások aránya [15]

Adathalmaz	**DNN**	**Fuzzy NN**
Breast cancer	0.23	0.25
Diabetes	0.26	0.28
King-Rook vs King-Pawn	0.06	0.07
Vote	0.05	0.29

3. táblázat. Optimális logikai kifejezés [15]

Adathalmaz	**Logikai kifejezés**
Breast cancer	$((28)uni(34))uni((6)uni(34))$
Diabetes	$1-((1)uni(6))$
King-Rook vs King-Pawn	$((9)uni(34))uni((22)uni(34))$
Vote	$((25)uni(31))uni((11)uni(37))$

urális hálót (DNN) használtak. A logikai kifejezésekben szereplő *uni* művelet az uninorma tulajdonságú kompenzáló operátort jelöli, azaz $S(x+y-0.5)$. A logikai kifejezés kiértékelése után a végeredmény egy folytonos érték 0 és 1 között. A bináris klasszifikációk esetén a vágóérték 0.5. A hibás besorolások aránya (misclassification rate) a legtöbb esetben hasonló (2. táblázat), viszont értelmezhetőség szempontjából a folytonos logikát használó neurális háló előnyt élvez. Eredményül a klasszifikációs érték mellett a magyarázó változók közötti optimális logikai kapcsolatot is visszaadja (3. táblázat).

5. Uninormán alapuló folytonos értékű neurális háló finomítása regularizációval

A fent bemutatott modell egy lehetséges továbbfejlesztése [18], a neurális háló regularizációs paraméterének optimalizálása.

A gépi tanulásban az optimalizáció kulcsfontosságú a modellek fejlesztéséhez. A regularizációs technika segít abban, hogy az optimalizációs folyamat során olyan modelleket válasszunk ki, amelyek kedvezőbb tulajdonságokkal rendelkeznek [19]. A Ridge regresszió L2 regularizációt használ, amely a modell együtthatóinak négyzetösszegével arányosan bünteti az optimalizálás során minimalizálandó értéket, emiatt előnyben részesíti a kevésbé extrém értékű együtthatókkal rendelkező modelleket, melynek következtében hatékonyan megelőzi a túltanulást, különösen

4. táblázat. Hagyományos neurális háló [18]

Adathalmaz	**Átlagos hibaarány**
Badges2	0.022
Breast cancer	0.284
Diabetes	0.270

akkor, amikor korlátozott mennyiségű tanuló adat áll rendelkezésre. Egy másik népszerű regularizációs módszer a Lasso, mely L1 regularizációt használ, ami a modell együtthatóinak abszolút összegével büntet. Ennek következtében előnyt élveznek a ritka, többnyire nulla együtthatókkal rendelkező modellek.

[18]-ban az a cél, hogy a szerzők az ember által is érthető, átlátható logikai kifejezést eredményező modellt fejlesszenek, ezért a ritka együtthatók kifejezetten előnyösek. Így ebben az implementációban az L1 regularizációt alkalmazzák.

A λ regularizációs együttható korlátozza az optimalizálás során hozzáadott büntetés mértékét. Ezt az együtthatót figyelmesen kell meghatározni, hogy meglegyen a megfelelő egyensúly a kevésbé komplex, de nagyobb pontosságú modellek között. Ha λ túl kicsi, akkor az eredményül kapott logikai kifejezés túl sok összetevőből áll, ha pedig túl magas értékű, akkor a modell várhatóan nem illeszkedik elég jól a tanuló adathalmazra, ezáltal hibás predikciókat eredményezve.

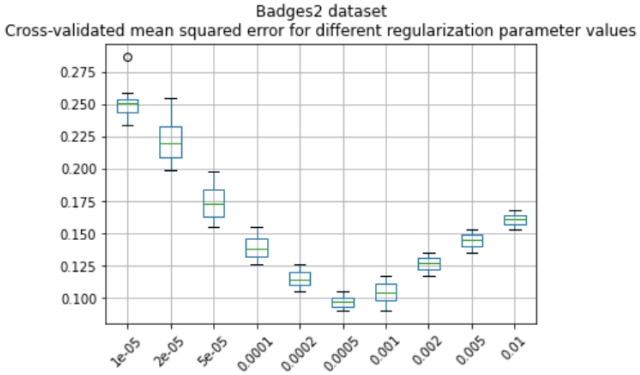
Az optimalizálás során mindegyik vizsgált regularizációs együttható ($\{1, 2, 5\} \times 10^{-5}$, $\{1, 2, 5\} \times 10^{-4}$, $\{1, 2, 5\} \times 10^{-3}$ és 10^{-2}) esetén 50 különböző futtatás történt. Az eredmények kiértékeléséhez a keresztvalidáció módszerét használták a szerzők [18].

A referenciaként használt mély neurális háló keresztvalidációval kiértékelt átlagos hibaarányai a 4. táblázatban találhatóak. A 2., 3. és 4. ábrán a három vizsgált adathalmaz hibaarányainak eloszlása látható különböző regularizációs paraméterekkel történő kiértékelések esetén. Látható, hogy az adott adathalmazhoz tartozó optimális regularizációs paraméter esetén a modell hasonló pontossággal teljesít, mint a referenciául szolgáló mély neurális háló, viszont emellett egy értelmezhető logikai kifejezést is ad eredményül.

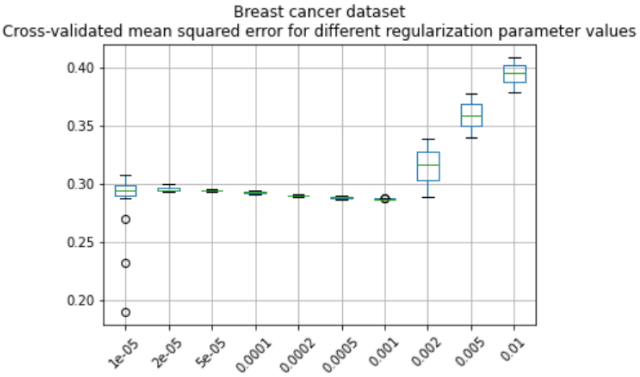
A különböző regularizációs együtthatókat keresztvalidációval kiértékelve a szerzők azt az eredményt kapták, hogy a kétváltozós klasszifikációs problémák esetén különböző adathalmazokra különböző regularizációs együttható lesz optimális. A számítások alapján optimális regularizációs paramétert alkalmazva rövid és könnyebben érthető logikai kifejezést adott eredményül a neurális háló, mint az eredeti, rögzített regularizációs paraméterrel.

A legjobban közelítő logikai kifejezéseket vizsgálva azt az eredményt kapták, hogy a leggyakrabban előforduló kifejezésekben szereplő művelet az *uni*, azaz az uninorma tulajdonságú kompenzáló operátor. Ezen logikai kifejezéseket, a

2. ábra. Keresztvalidáció a Badges2 adathalmaz esetén [18]



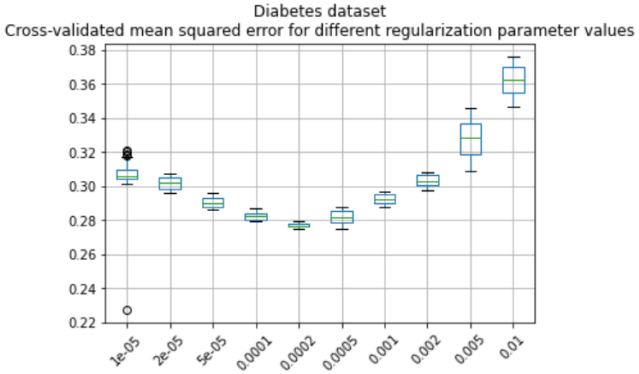
3. ábra. Keresztvalidáció a Breast cancer adathalmaz esetén [18]



keresztvalidáció során történő előfordulás darabszámát, a hozzá tartozó hibaarány értékeket és az előfordulásokhoz tartozó regularizációs paramétereket a 5. táblázat tartalmazza.

Az uninorma tulajdonságú kompenzáló operátor tapasztalt népszerűsége a hagyományos „és” és „vagy” operátorokkal szemben többek között azzal magyarázható, hogy a modell nem különbözteti meg a bináris klasszifikáció kétféle kérdésfeltevését. Például, a breast cancer, azaz a mellrák adathalmaz esetén kétféle kimenet lehetséges egy adott páciens vizsgálata során: vagy jelen van a betegség (1-es számmal jelölve) vagy nincs (0-val jelölve). Viszont, ugyanezt a problémát úgy is meg lehet fogalmazni, hogy mutassuk meg, mely páciensek esetén *nincs* jelen a betegség, azaz az egészségeseket jelöljük 1-essel, míg a betegeket 0-val.

4. ábra. Keresztvalidáció a Diabetes adathalmaz esetén [18]



5. táblázat. Leggyakrabban előforduló logikai kifejezések

Adathalmaz	Kifejezés	Hibaarány	Regularizációs együtthatók
Badges2	$(3)uni((3)uni(6))$	0.127	$\{1,2,5\}\times 10^{-4}$
Breast cancer	$((20)uni(29))uni((31)uni(32))$	0.292	$\{1,2\}\times 10^{-4}$
Diabetes	$(1)notuni(7)$	0.299	$\{1,2,5\}\times 10^{-5}, \{1,2\}\times 10^{-4}$

A logikai kifejezés kiértékelése során egy 0 és 1 közötti z számot kapunk, melynek függvényében a páciens diagnózisa 1, ha $z > 0.5$ vagy 0, ha $z < 0.5$. A neurális hálónak bármelyik típusú megfogalmazás esetén a megfelelő helyes választ kéne adnia, ami az egyik esetben z, a másikban viszont $1 - z$.

A 4. fejezetben bemutatott paraméteres aktivációs függvényt használó neurális háló többek között az $S(x+y-\nu)$ függvényben található ν paramétert optimalizálja, amely utána meghatározza az x és y értékeket összekapcsoló logikai operátort. [20] -ban a szerzők megmutatták, hogy az $S(x+y-\nu) = z$ egyenletbe behelyettesítve az ellentétesen megfogalmazott problémát, $S(x'+y'-\nu) = z'$, ahol $x' = 1-x$, $y' = 1-y$ és $z' = 1-z$, azt kapjuk eredményül, hogy $\nu = 0.5$. Ez a megközelítés egy lehetséges magyarázatot ad arra, hogy miért részesíti előnyben a neurális háló az *uni*, azaz az uninorma tulajdonságoló átlagoló operátort a diszjunkcióval és a konjunkcióval szemben.

6. Összefoglalás

Az értelmezhetőség egyre fontosabb tényezővé válik a gépi tanulásban. A hagyományos neurális hálók gyakran túl sok paramétert tartalmaznak, ami megnehezíti, hogy megértsük, miért hoznak bizonyos döntéseket. Az értelmezhetőség

javítása érdekében a folytonos logikai rendszerek és neurális hálók kombinációja egyre nagyobb figyelmet kap. Ezek az új modellek rugalmasságot és pontosságot kölcsönöznek a neurális hálóknak, miközben lehetővé teszik a jobb értelmezhetőséget és átláthatóságot. A folytonos logika segít megfogalmazni a neurális hálózatokban lévő transzformációkat, és ezáltal lehetőséget nyújt az eredmények természetes nyelvi interpretációjára. A folytonos logika és neurális hálók kombinálása olyan eredményekhez vezet, amelyek mind a hagyományos neurális hálók rugalmasságát, mind a folytonos logika értelmezhetőségét ötvözik. Ez az összekapcsolás egy ígéretes irány a gépi tanulásban és az értelmezhető modellek létrehozásában. Ennek a hibridizációnak egy ígéretes módja a vizsgált paraméteres aktivációs függvény, amely a paraméter választásától függően különböző, az érvelésben szereplő operátorokat tud modellezni. A paraméterek optimalizációjával lehetővé válik a neurális hálózat számára, hogy megtalálja a bemeneti változók közötti kapcsolatokat.

Köszönetnyilvánítás

A 2019-2.1.11-TÉT-2020-00217 számú projekt a Nemzeti Kutatási, Fejlesztési és Innovációs Alap támogatásával valósult meg, a 2019-2.1.11-TÉT-2020-00217 támogatási konstrukció keretében.

Hivatkozások

[1] J.-S.R. Jang et al. *Fuzzy modeling using generalized neural networks and kalman filter algorithm*, AAAI, Vol. **91**, pp. 762–767 (1991). DOI: 10.5555/1865756.1865795

[2] J.-S.R. Jang: *Anfis: adaptive-network-based fuzzy inference system*, IEEE Transactions on Systems, Man, and Cybernetics, Vol. **23** No. **3**, pp. 665–685 (1993). DOI: 10.1109/21.256541

[3] C.T. Lin and C.S. G. Lee: *Neural fuzzy systems: a neuro-fuzzy synergism to intelligent systems*, IEEE Transactions on Neural Networks, Prentice-Hall, Inc., Vol. **7** No. **5**, p. 1316 (1996). DOI: 10.1109/TNN.1996.536328

[4] S.X. Chen, H.B. Gooi, and M.Q. Wang: *Solar radiation forecast based on fuzzy logic and neural networks*, Renewable Energy, Vol. **60**, pp. 195–201 (2013). DOI: 10.1016/j.renene.2013.05.011

[5] C.L.P. Chen, Y.J. Liu, and G.X. Wen: *Fuzzy neural network-based adaptive control for a class of uncertain nonlinear stochastic systems*, IEEE Transactions on Cybernetics, Vol. **44** No. **5**, pp. 583–593 (2014). DOI: 10.1109/TCYB.2013.2262935

[6] E. Kayacan, E. Kayacan, and M.A. Khanesar: *Identification of nonlinear dynamic systems using type-2 fuzzy neural networks – a novel learning algorithm and a comparative study*, IEEE Transactions on Industrial Electronics, Vol. **62** No. **3**, pp. 1716–1724 (2015). DOI: 10.1109/TIE.2014.2345353

[7] J. Dombi and O. Csiszár: *Explainable Neural Networks Based on Fuzzy Logic and Multi-criteria Decision Tools*, Springer Nature, (2021). DOI: 10.1007/978-3-030-72280-7

[8] J. Dombi and O. Csiszár: *The general nilpotent operator system*, Fuzzy Sets and Systems, Vol. **261**, pp.1–19 (2015). DOI: 10.1016/j.fss.2014.05.011

[9] O. Csiszár, G. Csiszár, and J. Dombi: *Interpretable neural networks based on continuous-valued logic and multicriterion decision operators*, Knowledge-Based Systems, Vol. **199**, 105972 (2020). DOI: 10.1016/j.knosys.2020.105972

[10] T.R. Besold, A. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.U. Kühnberger, L.C. Lamb, P.M. Vieira Lima, L. de Penning, G. Pinkas, H. Poon and G. Zaverucha: *Neural-Symbolic Learning and Reasoning: A Survey and Interpretation*, Neuro-Symbolic Artificial Intelligence: The State of the Art, Vol. **342**, pp. 1–51 (2021). DOI: 10.3233/FAIA210348

[11] J. Dombi and Zs. Gera: *The approximation of piecewise linear membership functions and Łukasiewicz operators*, Fuzzy Sets and Systems, Vol. **154**, pp. 275–286 (2005). DOI: 10.1016/j.fss.2005.02.016

[12] J. Dombi and O. Csiszár: *Squashing Functions. In: Explainable Neural Networks Based on Fuzzy Logic and Multi-criteria Decision Tools*, Studies in Fuzziness and Soft Computing, Springer, Cham., Vol. **408**, pp. 121-134 (2021). DOI: 10.1007/978-3-030-72280-7_7

[13] D. Zeltner, B. Schmid, G. Csiszár and O. Csiszár: *Squashing activation functions in benchmark tests: Towards a more eXplainable Artificial Intelligence using continuous-valued logic*, Knowledge-Based Systems, Vol. **218**, 106779 (2021). ISSN 0950-7051, DOI: 10.1016/j.knosys.2021.106779

[14] L. Godfrey and M. Gashler: *A parameterized activation function for learning fuzzy logic operations in deep neural networks*, IEEE International Conference on Systems, Man, and Cybernetics (SMC), (2017). DOI: 10.1109/SMC.2017.8122696

[15] O. Csiszár, L.S. Pusztaházi, L. Dénes-Fazakas, M.S. Gashler, V. Kreinovich, and G. Csiszár: *Uninorm-like parametric activation functions for human-understandable neural models*, Vol. **260**, 110095 (2022). DOI: 10.1016/j.knosys.2022.110095

[16] K. Alvarez, J.C. Urenda, O. Csiszár, G. Csiszár, J. Dombi, G. Eigner, and V. Kreinovich: *Towards Fast and Understandable Computations: Which „And"- and „Or"-Operations Can Be Represented by the Fastest (i.e., 1-Layer) Neural Networks? Which Activations Functions Allow Such Representations?*, Acta Polytechnica Hungarica, Vol. **18** No. **2**, pp. 27–45 (2021). DOI: 10.12700/APH.18.2.2021.2.2

[17] R.R. Yager and A. Rybalov: *Uninorm aggregation operators.* Fuzzy Sets and Systems, Fuzzy Modeling, Vol. **80** No. **1**, pp. 111–120 (1996). DOI: 10.1016/0165-0114(95)00133-6

[18] L.S. Pusztaházi, G. Csiszár, M.S. Gashler, and O. Csiszár: *Parametric activation functions modelling fuzzy connectives for better explainability of neural models*, 2022 IEEE 20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY), pp. 77–82 (2022). DOI: 10.1109/SISY56759.2022.10036318

[19] S.L. Brunton and J.N. Kutz: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, Cambridge University Press, USA, 1st edition, (2019). DOI: 10.1017/9781108380690

[20] O. Csiszár, G. Csiszár, M. Ceberio, and V. Kreinovich: *Selecting the Most Adequate Fuzzy Operation for Explainable AI: Empirical Fact and Its Possible Theoretical Explanation*, (2023). CorpusID:259830061

Pusztaházi Luca Sára 1995-ben született Budapesten. BSc és MSc tanulmányait a Budapesti Műszaki és Gazdaságtudományi Egyetem Természettudományi Karának Alkalmazott matematika szakán sztochasztika szakirányon végezte. Diplomamunkáinak témája az internetes adatforgalom multifraktál analízise volt. A Nokiánál Data Scientistként dolgozott, melynek során telekommunikációs adatok elemzésével, predikciójával foglalkozott. Jelenleg másodéves PhD hallgató az Óbudai Egyetem Alkalmazott Informatikai és Alkalmazott Matematikai Doktori Iskolájában. Kutatási témái a megmagyarázható mesterséges intelligencia, a folytonos logika és a neurális hálók ötvözése. A PhD témakörében eddig 2 nemzetközi publikációja jelent meg.

PUSZTAHÁZI LUCA SÁRA
Óbudai Egyetem
pusztahazi.luca@uni-obuda.hu

Dr. Eigner György az Óbudai Egyetemen szerzett mechatronikai mérnöki alapdiplomát cum laude, a Budapesti Műszaki és Gazdaságtudományi Egyetemen pedig egészségügyi mesterdiplomát summa cum laude minősítéssel. Doktori fokozatát az Óbudai Egyetemen szerezte summa cum laude minősítéssel alkalmazott informatika tudományágban 2017-ben. Dr. Eigner György az Óbudai Egyetem Neumann János Informatikai Karának dékánja, valamint a Biomatika és Alkalmazott Mesterséges Intelligencia Intézet vezetője, ahol jelenleg egyetemi docensként tevékenykedik. Kutatási területe az élettani kapcsolatokra vonatkozó fejlett szabályozási módszerek, a biomedikai mérnöki tudomány, az ember-beavatkozással működő rendszerek, valamint a mesterséges intelligenciára épülő kibermedikai rendszerek.

EIGNER GYÖRGY
Óbudai Egyetem
eigner.gyorgy@uni-obuda.hu

Dr. Csiszár Orsolya az Eötvös Loránd Tudományegyetemen szerzett MSc diplomát matematikából, majd 2016-ban summa cum laude minősítéssel doktorált matematikából és számítástudományból az Óbudai Egyetemen. A németországi Aaleni Egyetemen a matematika és az alkalmazott mesterséges intelligencia professzora. Fő kutatási területe a megmagyarázható gépi tanulási módszerek fejlesztése.

CSISZÁR ORSOLYA
csiszar.orsolya@uni-obuda.hu
orsolya.csiszar@hs-aalen.de

PARAMETRIC ACTIVATION FUNCTION AND ITS SIGNIFICANCE IN THE INTERPRETABILITY OF NEURAL NETWORKS

Luca Sára Pusztaházi, György Eigner, Orsolya Csiszár

Artificial intelligence, especially deep learning models, are revolutionizing the business and technology world. One of the biggest challenges of deep learning today is solving the problem of interpretability. There is an increasingly pressing need to improve the transparency, performance, and safety of models (XAI: eXplainable Artificial Intelligence). Combining neural networks with continuous logic and multi-criteria decision-making tools can contribute to better interpretability, transparency, and safety in medical, engineering, and business applications. This approach, along with other emerging methods, belongs to neuro-symbolic hybrid artificial intelligence, a novel area of AI research that combines traditional rule-based approaches with modern deep learning techniques. Neuro-symbolic models have been proven to achieve high accuracy with significantly less data compared to traditional models. Neural networks and symbolic systems can complement each other's strengths and weaknesses, leading to precise, sample-efficient, and interpretable systems. This can improve decision-making, build trust in machine learning models, and lead to more efficient and effective processes across various industries. In this summary article, the latest results of our research group are presented in this field, comparing them to the most significant international trends.

Keywords: eXplainable AI, neural networks, fuzzy logic.

Mathematics Subject Classification (2000): 68T27.