**Hipotézis-tervezés PhD-disszertációkhoz - Konzisztens gépi tanuló modellezés beltéri felhasználói lokalizáció meghatározásának pontosítására**

Barta Gergő, Dr. Pitlik László

e-mail: Barta.Gergo@phd.uni-szie.hu, pitlik@my-x.hu

**Absztrakt**

A dokumentum célja, hogy a szinte senki által nem praktizált, de mégis minőségbiztosítási szempontból minimum-elvárásként értelmezhető eljáráshoz, vagyis a kérdőívek előzetes véletlenszám-alapú kiértékeléséhez hasonlóan egy teljes boltívet jelentő gondolatkísérletet vázoljon fel egy valóban PhD-értékű hipotézis bizonyítására, ahol létezik valódi és objektív verseny (benchmark), ill. létezik stratégiai (absztrakciós/komplexitás szintet érintő) különbség az alternatív megoldások között. Valamint, ahol a jó fogalma több mint egy tényezős és a megoldási alternatívák száma is legalább kettő. Tehát, ahol nem egyszerű szélsőértékkeresés a legjobb megoldás meghatározása. Összefoglalóan, a cél a lehető legrövidebb, már releváns szöveg megalkotása. A cikkben közölt alkalmazott kutatás fókusza beltéri felhasználók helymeghatározásának modellezése és predikciója volt, mely egy korábbi kutatás folytatásaként kísérelte meg a rendelkezésre álló adathalmaz alapján konzisztens módon a modellpontosság növelését.

**Kulcsszavak**

hermeneutika, automatizálás, nem-tudom-rendszerválasz, konzisztencia-alapú context free és kontextus-függő modellezés

**Abstract**

The goal of the paper is to demonstrate how a potential hypothesis can be articulated and evaluated for a real PhD-degree. The process is simple: a whole thinking experiment is needed where it is possible to build a benchmark (objective competition between alternative solutions). Parallel, it is relevant to be capable of defining a scale for the goodness of models. The focus of the research of this article was to model and predict the location of indoor users, which, is the continuation of a previous research, attempted to consistently increase model accuracy based on the available data set.

**Keywords**

hermeneutics, automation, none-system-answer, consistence-based context-free and/or context-depending modelling

**Biográfiai megjegyzés**

A cikk az eredetileg megjelent „Implementing and Evaluating Different Machine Learning Algorithms to Predict User Localization by the Strength of User Devices’ Wi-Fi Signal” (Barta, 2018) angol nyelvű cikk folytatása, melyet hasonlóan ezen dokumentum egyik szerzője készített.

Az eredeti adathalmazhoz az alábbi referencia tartozik:

Lichman, M. (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Az eredeti adathalmaz az alábbi linken tölthető le: [https://archive.ics.uci.edu/ml/datasets/Wireless+Indoor+Localization](https://archive.ics.uci.edu/ml/datasets/Wireless%2BIndoor%2BLocalization)

Az alábbi cikk hasonlóan a hivatkozott adatbázist dolgozza fel:

Jayant, G, Rohra, Boominathan, Perumal, Swathi, Jamjala Narayanan, Priya, Thakur, and Rajen B Bhatt (2017): User Localization in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithm with Neural Networks. In Proceedings of Sixth International Conference on Soft Computing for Problem Solving, pp. 286-295.

Előzmény cikknek tekinthető még:

Pitlik, M, Pitlik, L. (2018): Konzisztencia-orientált modell-hibridizáció látszólag triviális műszaki jelenségek kezelésére. MAGYAR INTERNETES AGRÁRINFORMATIKAI ÚJSÁG : 239 pp. 1-15. Letölthető: <http://miau.my-x.hu/miau/239/konzisztencia_alapu_hibridizacio_v1.docx>

**Bevezetés**

Vegyünk egy vélhetően egyszintes épületet, mely négy irodát tartalmaz (1-2-3-4). Vegyünk 7 wifi-routert (1-2-3-4-5-6-7), melyek bárhol lehetnek, akár egy irodában több is. Az épületben mozgó mobiltelefon és a routerek közötti jel erőssége legyen adott (X1-X2-X3-X4-X5-X6-X7) az érintett iroda számával (Y1) együtt 1600 tanulási esetre és 400 teszt esetre. A feladat, a 7 jelerősség alapján a szobaszám megbecslése minél pontosabban.

**Célkitűzések**

1. Megtalálni a közelíthetőség maximumát, azaz megtalálni azt a pontosságra vonatkozó eredményt, mely felé vélhetően semelyik modell sem tud menni. Mivel az adathalmazban nincs antagonizumus (nincs két olyan egyenlő adatsor (X1-2-3-4-5-6-7&Y), melyhez különböző célváltozó tartozik), ezért az elméleti pontosságnak 100%-nak kell lennie, annak figyelembe vétele nélkül, hogy egy adott modell 100%-os pontosságnál túlilleszkedést produkálhat. Bár megjegyzendő, hogy egy tévesen rögzített/mért X vagy Y érték is vezethet tökéletlen becsléshez (vö. összeadó automata, mely két inputszám összegét adja outputként – ha a tripletek bármely eleme sérül, akkor az automata működését feltárni akaró modell sem lesz tökéletes). Mivel előzetesen nem ismert az a tény, hogy a mért adatok 100%-osan tökéletesen lettek mérve, ezért azzal a kezdeti feltételezéssel élünk, hogy minden adatsor tökéletesen rögzített, tehát nincs tévesen mért adat az adattáblában.
2. Olyan konzisztens (logikai bizonyításokon keresztül objektíven magyarázható, nem-véletlenszerű) modell építése volt, mely a referált adathalmaz egy előre meghatározott részhalmazán (ezentúl: „teszthalmaz”) képes legalább a közelíthetőségi maximumot, de jobb esetben a 100%-ot elérni, azaz képes bármelyik tesztadatponthoz tartozó osztályt helyesen megbecsülni. Az eredeti cikkben (Barta, 2018) elért legmagasabb pontosság egy egyszerű modellel 97.67% volt, melyet a KNN algoritmussal sikerült elérni. Hibrid modellel a legjobb teljesítmény 98.33% volt, mely 6 modell együttes alkalmazását jelentette. A teszthalmaz összesen 600, míg a tanítóhalmaz (az a halmaz, amelyen az algoritmusok az összefüggéseket tanulták) 1400 adatpontot tartalmaz.
3. Mivel egy modell performanciáját az adatok minősége és az elérhető információ nagyban meghatározza, ezért jogosan felmerül az igény, hogy az elérhető adatokból van-e lehetőség több információ birtokába jutni valamilyen matematikai származtatással. Tehát a célkitűzés a maximális magyarázhatóság feltárása nem kizárólag a modellek hibridizációjával, hanem új adatok létrehozásával a meglévő adathalmazból.
4. Mennyire lehet a tanulást redukálni a teszt-sikeresség quasi szinten tartása mellett?
5. A biográfiai részben ismertetett cikk (Jayant et al., 2017) által publikált eredmények megugrása. A publikációban a szerzők egy szűkített adatbázissal dolgoztak, ahol kizárólag 3 iroda adatait vették számításba. A cikkből a tréning- és tesztadatok számára nem derül fény, így az eredmények összehasonlítása problémába ütközik. 6 különböző algoritmussal az alábbi teljesítményt érték el:
6. táblázat: Jayant et al. (2017) által publikált teszt-eredmények.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algoritmus | PSO-NN | GSA-NN | PSOGSA-NN | FPSOGSA-NN | SVM | Naive Bayes |
| Performancia (%) | 64.66 | 77.53 | 83.28 | 95.16 | 92.68 | 90.47 |

**Hipotézisek**

H1: Az NN7 tanulási pontosságát az NN6 modellek megverhetik úgy a tesztben, mint a tanulásban, akár egyetlen egy NN6-modell esetén is – vagyis a több adat nem biztos, hogy jobb modell a hagyományos, azaz nem konzisztencia-alapú modellezési gyakorlat keretei között, ami, ha igaz, akkor a konzisztencia-orientált modellezés kötelező!

H2: Ha minden NN6-model azonos szobaszámot tippel, akkor nincs ok és mód ettől eltérőt vélelmezni, így a NN6-becslések szórása, ha nem nulla, akkor ez a részhalmaz tekinthető egy context-free konzisztencia-kritériumnak, melytől elvárjuk, hogy az NN7-alapú 3 jóság-kritérium mindegyikét meghaladja a nem-tudom-részhalmaz leválasztása utáni becslések jósága – sőt, az NN7&NN6 halmaz minden jóság maximuma is meghaladható ezen elven.

H3: Ha csak az NN6\_1&4 és az NN6\_6&7 magas pár-korrelációjú modellek esetére várjuk el kontextus-függő korrelációként azt, hogy ezek páronként azonos becslést nem eredményező szűrlete legyen a nem-tudom-rendszerválasz feltétele, akkor elvárható, hogy a tanulás-sikeresség, a teszt-sikeresség, a teszt/tanulás-arány mellett a nem-tudom-részhalmaz minimalizálása részlegesen/teljesen jobb eredményre vezet.

**Módszertan**

A korábbi hivatkozott cikkben közölt pontossági mutató javítása konzisztens módon, az adathalmazban rejlő logikai kapcsolatok feltárásával történt. A rendelkezésre álló alaphalmazt egyváltozós elemzésekkel vizsgálva kiszámításra került osztályonként az átlag, szórás, minimum és maximum, továbbá a többváltozós Pearson-korreláció, azzal a céllal, hogy az előre nem ismert router- elhelyezkedésekről további többlet információt lehessen szerezni. Ezzel párhuzamosan főkomponens-analízissel a kezdeti változók transzformálása történt meg, melyben két különböző változóba történt az adatok tömörítése, azért, hogy azokat egy kétdimenziós térben ábrázolni lehessen, ezzel vizualizálni tudjuk egyrészt a felhasználók adott pontban ismert lokalizációját, másrészt a routerek elhelyezkedését. A vizualizáció célja az adathalmaz mélyebb megismerése volt, feltételezve, hogy ismerve a routerek elhelyezkedését további kapcsolatokat (egészen pontosan azt, hogy a két-két legközelebbi routerhez köthető válaszpárok azonosak illene, hogy legyenek, mert miért is lennének mások – s itt lép be a 100%-ot beárnyékoló erőtér: vajon a pár-azonossági előszűrés azért vezet hibára, mert rossz a pár-azonossági elvárás, vagy mert pl. hibásak az adatok?) lehessen az adatok között felfedezni. A vizualizációt követően ismertté vált a routerek elhelyezkedése, ezáltal lehetővé vált, hogy újra gondoljuk, hogy miként lehet az eredeti adathalmazt megváltoztatni, ahhoz, hogy a modellek pontosságán javítsunk. Új adathalmaz felhasználása után neurális háló modellillesztése történt, majd az eredmények összevetése a korábbi cikk eredményeivel.

A routerek 2D-s vizualizációját ugyan nem támogatja, de a közelségüket a korreláció-számítás is leleplezi, sőt, az 1. és 3.és 4. táblázat átlagai, maximuma és/vagy minimuma alapján is hasonló pár-képzés valószínűsíthető lényegesen kevesebb számítási igénnyel, mint pl. a főkomponens elemzés, vagy más pl. kombinatorikai sík-letapogatás.

**Eredmények**

Az osztályonkénti és változónkénti egyváltozós elemzések eredményeit az alábbi táblázatok szolgáltatják.

1. táblázat: A tréning-halmazra vonatkozó egyváltozós statisztikák (átlag)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **Y** |
| -62.42 | -56.37 | -60.48 | -64.08 | -70.23 | -82.90 | -84.00 | 1 |
| -36.74 | -56.09 | -55.91 | -37.75 | -67.54 | -72.33 | -73.41 | 2 |
| -49.82 | -54.94 | -52.83 | -50.67 | -63.30 | -81.39 | -82.36 | 3 |
| -60.22 | -55.29 | -50.73 | -61.39 | -49.39 | -87.11 | -87.01 | 4 |

1. táblázat: A tréning-halmazra vonatkozó egyváltozós statisztikák (szórás)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **Y** |
| 3.36 | 3.24 | 3.76 | 3.59 | 4.62 | 3.69 | 4.04 | 1 |
| 8.75 | 3.35 | 4.12 | 8.03 | 5.28 | 4.52 | 4.74 | 2 |
| 2.76 | 3.79 | 3.17 | 3.62 | 3.58 | 3.61 | 4.18 | 3 |
| 3.02 | 3.24 | 4.19 | 3.81 | 3.54 | 3.42 | 3.48 | 4 |

1. táblázat: A tréning-halmazra vonatkozó egyváltozós statisztikák (minimum)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **Y** |
| -74 | -69 | -73 | -76 | -89 | -97 | -94 | 1 |
| -52 | -74 | -71 | -53 | -85 | -89 | -92 | 2 |
| -63 | -68 | -73 | -63 | -77 | -92 | -93 | 3 |
| -71 | -66 | -60 | -74 | -62 | -96 | -98 | 4 |

1. táblázat: A tréning-halmazra vonatkozó egyváltozós statisztikák (maximum)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **X5** | **X6** | **X7** | **Y** |
| -55 | -47 | -50 | -54 | -60 | -74 | -74 | 1 |
| -10 | -46 | -46 | -11 | -56 | -62 | -63 | 2 |
| -43 | -45 | -45 | -40 | -54 | -74 | -73 | 3 |
| -54 | -48 | -40 | -52 | -36 | -76 | -78 | 4 |

A változók közötti összefüggések kiszámítása Pearson-korrelációval történt.

1. táblázat: A tréning-halmazra vonatkozó Pearson-korreláció változónként

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
| X1 | 1.00 | 0.01 | 0.04 | 0.92 | -0.25 | 0.73 | 0.69 |
| X2 | 0.01 | 1.00 | 0.28 | 0.02 | 0.20 | 0.08 | 0.06 |
| X3 | 0.04 | 0.28 | 1.00 | 0.07 | 0.61 | -0.10 | -0.08 |
| X4 | 0.92 | 0.02 | 0.07 | 1.00 | -0.25 | 0.72 | 0.68 |
| X5 | -0.25 | 0.20 | 0.61 | -0.25 | 1.00 | -0.43 | -0.37 |
| X6 | 0.73 | 0.08 | -0.10 | 0.72 | -0.43 | 1.00 | 0.72 |
| X7 | 0.69 | 0.06 | -0.08 | 0.68 | -0.37 | 0.72 | 1.00 |

Az alábbi öt táblázat alapján a következő összefüggéseket lehet vélelmezni:

1. Az X1 és X4 között erős korreláció van (0.92), az egyváltozós statisztikák közötti átlagos különbségek a két változó között a legkisebb (más változókhoz viszonyítva), ezért vélhetően az X1 és X4 routerek közel lettek telepítve egymáshoz.
2. Az X2 és X3 egyváltozós statisztikák közötti átlagos különbségek a két változó között a legkisebb (más változókhoz viszonyítva), ezért vélhetően az X2 és X3 routerek közel lettek telepítve egymáshoz, azonban ezt a korrelációs mátrix nem támasztja alá.
3. Az X6 és X7 között jelentős korreláció van (0.72), az egyváltozós statisztikák közötti átlagos különbségek a két változó között a legkisebb (más változókhoz viszonyítva), ezért vélhetően az X6 és X7 routerek közel lettek telepítve egymáshoz.
4. Az X5-ös router az egyváltozós statisztikák alapján az X3-as routerhez állhat a legközelebb, ráadásul a kettő között jelentős (0.61) korreláció van, de ezt nem erősíti az átlag, max, min – mint a többi korreláció-pár értelmezését és az 2. ábra sem.
5. ábra: Benchmarkok és alternatív megoldások többrétegű jóság-kritériumrendszerrel 

A H1 kapcsán az NN6-X2 és NN6-X7 külön-külön egyedül is jobb találati arányt produkáltak a tanulásban és a tesztben is, sőt a teszt/tanulás arányban is (1. ábra bal felső szekció).

A H2 kapcsán az összes NN6 modell becslésazonosságát jelző szórás=0 elvárás teljesülése esetén az NN7&NN6 maximális tanulási, teszt és teszt/tanulás arányai is meghaladásra kerültek (vö. lila-fehér jelzések az 1. ábrán). A nem-tudom-rendszerválasz átlagosan 5.2% (vö. 1. ábra közepe).

A H3 kapcsán a csak az NN6\_1&4 és az NN6\_6&7 páronkénti becslésazonosságát szűrése, azaz kontextus-függő konzisztencia-kritériumként használva a tanulási, teszt találati arányok nem javultak, de a nem-tudom-rendszerválaszok aránya és a teszt/tanulás aránya javult, vagyis fennáll a versengő megoldások jóságkritériumainak antagonizmusa, azaz a lehet-e minden megoldás másként egyformán jó komplexitás-emelés kényszere a Jóság fogalmának modellezésénél (vö. 1. ábra jobbra alul).

Főkomponens-elemzéssel megtörtént az adathalmazra vonatkozóan a dimenziócsökkentés, azaz a kezdeti 7 változó által közölt információ 2 változóba lett tömörítve, úgy, hogy a kezdeti adathalmazt transzponáltuk. Kettő változó esetén az adódó variancia 5.8%, mely az információ-veszteségre utal. A változók értékei koordinátákként is felfoghatóak, ezért a 7 routert vizualizálni tudjuk, feltehetően 94.2%-os pontossággal, melyet az 2. ábra szemléltet.

1. ábra: Routerek vizualizációja



Az egyváltozós statisztikák, a Pearson-korreláció és a vizualizáció után egyértelmű válik, hogy elég evidenciával rendelkezünk ahhoz, hogy kijelenthető legyen, hogy az X1 és X4 router valóban egymáshoz közel lettek telepítve, hasonlóan az X6 és X7 routerekhez. Azonban a vizualizáció alapján az X2 és X3 routerek is párban lettek telepítve, melyet a statisztikai számítások kevésbé támasztanak alá. Az X5 router, ahogy korábban sejteni lehetett, magányosan helyezkedik el, vélhetően legközelebb az X2 és X3 routerhez áll a legközelebb. Ezek a vélelmek alternatív nem-tudom-rendszerválasz-szűrők!

Mindezek alapján racionálisan következtethető, hogy a párban elhelyezett routerek redundáns információt tartalmaznak, ezért a modellezés kapcsán 1-1 pár elhanyagolásával feltételezhető, hogy a modellezés pontossága nem romlik, azonban az X5 egy fontos változót jelent a modellezés szempontjából, mivel a router alapján történt mérések többletinformációt jelentenek.

Ugyanezen transzformációs logikával vizualizálható a felhasználók elhelyezkedése is. A főkomponens-elemzés ez esetben 30% varianciával járt.

1. ábra: Felhasználók vizualizációja



A vizualizáción azt vehetjük észre, hogy többé-kevésbé kirajzolódnak a szobák, melyekben a routerek lettek elhelyezve. Jó pár esetben észrevehető, hogy feltételes anomáliával találkozhatunk, amikor is egy-egy szobában feltűnik egy másik szoba színével jelzett felhasználó. Az egyes vizualizált pontok egymással történő átfedése magyarázatot ad arra, hogy miért nem volt lehetséges a 100%-os megoldás elérése. Az adathalmazban kizárólag a routerek által mért jelerősség adott, azonban a helymeghatározást számtalan más tényező is befolyásolhatja, mint pl. a routerek minősége, mobileszközök minősége, falak áteresztőképessége, felhasználók által viselt ruházat stb., így nem lehetünk minden információ 100%-os birtokában, feltételezhetően a 100%-os megoldás sem elérhető.

Ha az adathalmaz felhasználását újra gondoljuk a modellezéshez, ezért ésszerűnek tűnik az, ha kísérletet végzünk a párban álló routerek 1-1 elhagyásával. A 6 routerre programozott neurális háló teljesítményét a következő táblázat szemlélteti.

1. táblázat: Neurális hálók és konzisztencia-modellek parallel teljesítményei

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modell** | **Pontosság a tréning-halmazon****(minél nagyobb, annál jobb)** | **Pontosság a teszt-halmazon (minél nagyobb, annál jobb)** | **Nem-tudom tesztválaszok aránya** **(minél kisebb, annál jobb)** | **Teszt/tanulás arány** **(minél nagyobb, annál jobb)** |
| **A** | **B** | **C** | **D** | **E=C/B** |
| NN - 7 Router[[1]](#footnote-1) | 97.50% | 96.67% | 0% | 99.15% |
| NN - 6 router -X1 kihagyásával | 97.50% | 96.67% | 0% | 99.15% |
| NN - 6 router -X2 kihagyásával | 98.64% | 98.33% | 0% | 99.69% |
| NN - 6 router -X3 kihagyásával | 98.57% | 97.33% | 0% | 98.74% |
| NN - 6 router -X4 kihagyásával | 98.64% | 96.17% | 0% | 97.50% |
| NN - 6 router -X5 kihagyásával | 96.36% | 94.83% | 0% | 98.41% |
| NN - 6 router -X6 kihagyásával | 99.14% | 97.50% | 0% | 98.35% |
| NN - 6 router -X7 kihagyásával | 99.00% | 98.00% | 0% | 98.99% |
| LR (Barta, 2018) | n/a | 97.17% | 0% | n/a |
| SVM (Barta, 2018) | n/a | 97.50% | 0% | n/a |
| KNN (Barta, 2018) | n/a | 97.67% | 0% | n/a |
| DT (Barta, 2018) | n/a | 96.33% | 0% | n/a |
| NN (Barta, 2018) | n/a | 96.67% | 0% | n/a |
| K-means (Barta, 2018) | n/a | 92.83% | 0% | n/a |
| Hibrid modell I. (2018) | n/a | 98.17% | 0% | n/a |
| Hibrid modell II. (Barta, 2018) | n/a | 98.33% | 0% | n/a |
| Szórás=0 modell (2020)vö. 1. ábra | 1342/1345=99.78% | 549/551=99.64% | 8.17% | 99.86% |
| &&-modell (2020)vö. 1. ábra | 99.19% | 99.10% | 7.17% | 99.91% |
| 18-as kiegészítő modell (2020)vö. 1. és 4. ábra | 1354/1357=99.78% | 555/557=99.64% | 7.17% | 99.86% |

A táblázatból az alábbiak állapíthatók meg:

1. 6 routerrel történő modellezés esetén a tréning- és teszthalmazon a legrosszabb eredményt X5 kihagyásával értük el, ezért látható, hogy az X5, mint egyedülálló router, információtartalma valóban jelentős a performancia növelése érdekében. Az X5 speciális szerepét még nem használtuk ki eddig, vagyis új input-OAM lehetne egy X5-X(i), mely alapján új modellek készíthetők, melyekre a már alkalmazott lépések akár jobb eredményre is vezethetnek? Vagyis a minden adat értéke az X5-höz való viszonyától függ, egy kontextus-függő konzisztencia-réteg lehetne?!
2. 6 routerrel történt modellezés esetén mindegyik esetben (kivéve, ha az X5-t hagyjuk el) legalább annyira jó eredményt tudunk elérni, mintha 7 routerrel történt volna a modellalkotás.
3. 6 routerrel történt modellezés esetén az X5 és X4 kihagyásán kívül, legalább annyira jó eredményt tudunk elérni, mintha 7 routerrel történt volna a modellalkotás.
4. A legjobb eredményt a tréninghalmazon az a 6 routeres modell volt képes elérni, ahol az X6-t hagytuk ki, a performancia 99.14%.
5. A legjobb eredményt a teszthalmazon az a 6 routeres modell volt képes elérni, ahol az X2-t hagytuk ki, a performancia 98.33%, mely performancia megegyezik Barta (2018) által korábban közölt legjobb hibrid modellel (vö. 7. táblázat).

**Vita**

A context free konzisztencia-kritérium (szórás=0) azért univerzális, mert nem kell ennek használatához semmit feltételezni a problémáról magáról. Míg az 1&4 és 6&7 modellek pár-azonossága azon a probléma-alapú feltételezésen alapul, hogy a routerek közelsége a becslés azonosság elvárásával párosulhat mérnöki szempontból. Az itt bemutatott neurális háló modellek helyett/mellett bármilyen más alapmodellek is készíthetők (pl. döntési fák, regressziós modellek, hasonlóságelemzések, stb.). Az itt bemutatott konzisztencia-kritériumok mellett más context-free és kontextus-függő megoldások is készíthetők (pl. a routerek távolságát becslő geometriai háttérmodell alapján más rokonságok is vélelmezhetők, s eleve többféleképpen lehet 2+2 router-párazonosságot vélelmezni).

A nem-tudom-rendszerválaszok aránya antagonisztikusan hat a legjobb megoldás címért folyó „harcban”. Vagyis a hasznos alternatívák létének bizonyítása után megkerülhetetlen az alternatív megoldások anti-diszkriminatív (context-free) megversenyeztetése, ahol annál ideálisabb egy megoldás, minél kisebb a nem-tudom-rendszerválaszok aránya, minél nagyobb a tanulási siker, a tesztelési siker, a teszt/tanulás sikerarány, a teszt/tanulás objektumdarabszám-arány, stb.

**A modellek összehasonlítása**



1. Ábra: Összehasonlító OAM (forrás: 7. táblázat alapján)



1. Ábra: A többlépcsős összehasonlítás eredményei (forrás: <https://miau.my-x.hu/miau/263/wifi_plus.xlsx>)

A 4. ábra objektumai azok a modellek, melyek mind a 4 attribútuma a 7. táblázat alapján adott. Mint az antidiszkriminatív elemzések alapján jól látható, a nem-tudom-válaszok jelensége kisebbségi helyzetet idéz elő, vagyis a nem-tudom nullánál nagyobb aránya sajátos/önálló alcsoportként beáll a normaérték köré – quasi elnyomás alá kerül a nagyobb számosságú mindenkor válaszoló modell-alternatívákkal szemben, de az utolsó modell (18) így is norma fölé kerül.

Ha azonban (vö. 5. ábra) immár csak a 4. ábra norma (1000) feletti objektumai alapján folytatódik a vizsgálat, akkor elsőként a minden alternatíva másként egyforma filozofikus választ kapjuk eredményül, vagyis a robot (COCO-Y0: <https://miau.my-x.hu/cocoy0>) látszólag visszaadja a döntést a felhasználónak, hogy a nem-tudom-minőséget miként akarja értékelni. De a step-eljárás keretében a fel nem használt teszt/tanulás arány mégis átsegíti az elemzőt ezen a dilemmán és egyértelműsíti, hogy a legjobb modell a konzisztens (18) csúcsmodell!

**Konklúzió**

A 98.33%-os performanciát a tréninghalmazon elérve a korábbi cikk (Barta, 2018) összes egyszerű, nem hibridizált modelljét megverte, ezzel kijelenthető, hogy a konzisztencia (itt azt tekintjük konzisztencia-növelésnek, hogy feltételezni merjük, hogy egyes X-k kifejezetten feleslegesek, zavarók, akkor ez lehet akár egy context-free konzisztencia-kritérium is, mert bármilyen OAM esetén igaz lehet – vö. összeadó automata szabadsági fok) szellemében, az adathalmaz belső logikai struktúrájának feltárásával jobb eredményt tudtunk elérni, ugyanazon modellt felhasználva, mint a korábbi kísérletekben.

A hipotézisek mindegyike teljesült. A konzisztencia-alapú modellezés kötelező és hasznos jellege félreérthetetlenül bizonyítást nyert. A pontosság növelésére további lehetőségek is vélelmezhetők. További (PhD-címet érő) hipotézisek valószínűsíthetők.

A 2018-as és a 2020-as modellek összevetése kapcsán (l. 7. táblázat) elmondható, hogy a nem-tudom válaszok nem 0%-os arányát az egyre pontosabb komplex (konzisztencia-vezérelt) modellek kompenzálják, ahol az összes modell-jóság-attribútum alapján a lehet-e-minden-modell-másként-egyforma anti-diszkriminatív ellenőrzés a nem antagonista modell-variánsok versenyének győztesét képes levezetni.

**Az antagonisztikus attribútumkészlet alapján a nem-tudom opcióval élő és nem élő modellek összevethetők és létezik a nem-tudom-válaszok ekvivalenciája, vagyis az a mértékű javulás a kezelni akart részhalmazon, mely kompenzálni képes a nem-tudom-opció általi működés-korlátozást magát.**

**Jövőkép**

A tanulási minta hibás és a helyes becslései által produkált bináris jel (Y2) klasszifikáció alapú értelmezése egy univerzális hasonlóságelemzési almodellek döntési faként való integrálása keretében automatizáltan, context-free módon önmagában is új hipotézisek forrásaként értelmezhető.

**Hivatkozások**

Barta, Gergő (2018): Implementing and Evaluating Different Machine Learning Algorithms to Predict User Localization by the Strength of User Devices’ Wi-Fi Signal. Sefbis Journal 12, pp. 2-11.

Jayant, G, Rohra, Boominathan, Perumal, Swathi, Jamjala Narayanan, Priya, Thakur, and Rajen B Bhatt (2017): User Localization in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithm with Neural Networks. In Proceedings of Sixth International Conference on Soft Computing for Problem Solving, pp. 286-295.

Lichman, M. (2013): UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Pitlik, M, Pitlik, L. (2018): Konzisztencia-orientált modell-hibridizáció látszólag triviális műszaki jelenségek kezelésére. MAGYAR INTERNETES AGRÁRINFORMATIKAI ÚJSÁG : 239 pp. 1-15. Letölthető: <http://miau.my-x.hu/miau/239/konzisztencia_alapu_hibridizacio_v1.docx>

**Melléklet**

A fenti szórás és &&-típusú alakzatok direkt konzisztencia alakzatok, ráadásul a szórás=0 alakzat context free, míg az &&-típus kontextus függő.

Létezhetnek további kontextus-függő logikai alakzatok is, melyek a „nem-az” fogalmának bevezetését várják el.

H4: Létezhetnek hasznos szabályok „nem-az” jelleggel.

H5: Létezhetnek hasznos szabályok másodlagos, harmadlagos becslésekről a „nem-az” szabályok nyomán.

1. ábra: A szórás=1 és az &&-típus különbségét jelentő 18 eset értelmezése



H4: A 18 esetből 11 téves becslést ad, tehát kijelenthető, hogy ezen 18 esetben (melyek feltárása a szórás=1, sem-sem=2) többet nyerünk, ha ezen becslések ellenkezőjét fogadjuk el. Mivel azonban a 4 szoba nem bináris következmény, így elsődlegesen az ellenkező értelmezés nem direkt becslést ad, hanem egy mi nem-az értelmezést. Ami elsődlegesen egy új nem-tudom kategória is lehetne, de ezzel nem javulna egyelőre a pontosság, mert nem születik becslés, csak kizárás (mi-nem-lehet, mi-nem-az).

H5: Ha értelmezhető másodlagos becslés (pl. a NN6-alapú becslések közül a leggyakoribb), akkor a kizárás kiegészülhet direkt becsléssel. Sőt, harmadlagosan az is vizsgálható, hogy csak az a másodlagos becslés érvényesítendő, ahol az elsődleges becslés nem a leggyakoribb már önmagában is az NN6-halmazban.

Konklúzió: a H4 és H5 alapján a 18 eset a tanulásra érvényes, onnan feltárt szabályok alapján 100%-ban sikeres találatként értelmezhető, vagyis az 1. ábrán középen alul látható színes (zöld, piros) legjobb és benchmark értékek mellé egy a legjobbnál is jobb megoldás vezethető le a nem-az, a másodlagos és a harmadlagos becslések alapján. Vagyis 1914 esetből immár nem 16 hiba van, hanem továbbra is csak 5 hiba, mint korábban a 1896 esetből, mert mind a 18 eset (1914-1896=18) helyesen klasszifikálható a H4 és H5 felhasználásával, azaz további alapvetően kontextus függő szabályrészletekre alapozva. A kontextus-függőség azért áll fenn, mert az NN6-sorozat maga kontextus-függő, míg a szórás=0 elvárás bármilyen alternatív modellezés esetén is lehetséges elv lenne, azaz context free. Hacsak nem tekintjük univerzálisnak magát a felismerést, hogy n-1 változóval képzett modellek legyenek mindenkor az alternatív modellek – ami automatikusan minden fenti konzisztencia-alakzatot context-free-ként enged értelmezni.

Ez a gondolatsor az irodalmi előzményekben tárgyalt kétpólusú részmodellek logikájának egy speciális eseteként is felfogható.

H6: A nem-tudom-ágon önálló modellezési feladat indítható

2000-1914=86 eset marad a szórás=0 elv nyomán a nem-tudom ágra. Háttérszámítások: [https://miau.my-x.hu/miau/262/wifi\_y=1.xlsx](https://miau.my-x.hu/miau/262/wifi_y%3D1.xlsx)

A 86 esetet feldolgozó önálló tanulási feladatból 43 eset volt tanulási eset és 43 teszteset. A 7-változós modell összes találata a 86 esetre 54 volt, ezen belül a tanulásra 25/43. Ez önálló tanulási feladatként 42/43-ra emelkedett, ahol a tesztsikeresség 30/43-ra nőtt egy additív induktív szakértői rendszer-modell alapján. Ez annyiban erősíti a konzisztencia-elvek mentén való találati arány maximalizálást, hogy önálló feladatokat lehet definiálni konzisztencia-elvek mentén kijelölve az önállóként értelmezendő tanulási minták részleteit. A 25-ről 42-re növekvő találati arány világosan jelzi, hogy az egyéb módon értelmezhető tanulási mintáktól leválasztott kérdés a kisebb mennyiség és a nagyobb egyediség/önazonosság révén is jobban tanulható, ami azonban a konkrét esetben az eleve ritka tanulásénál jobb tesztsikeresség meghaladására alig volt képes, de képes volt…

A nyers (eredeti router-adatok) alapján a 86 elemű tanulási minta 31/43 tesztsikerességgel értelmezhető (32/43 tanulási sikeresség mellett), ami egy kiegyensúlyozottabb tanulás-teszt viszont jelent, mint a modell-eredmények alapján.

A döntési fák alapvetően nyers-adatokból építenek logikai láncokat tisztán a találati arányok maximalizálása mentén, míg a konzisztencia-orientált modellezés a modell-részeredmények kapcsolatából vezet le újabb és újabb tanulási részfeladatokat (vö. 86-elemű minta).

1. A 7 routeres modell performanciája Barta (2018) cikke alapján történt bemásolásra. [↑](#footnote-ref-1)