Production Systems and Information Engineering
Volume 11 (2), pp. 73-87
doi: 10.32968/psaie.2023.2.7

ALKALMAZÁSPÉLDÁK A HFRIQ-LEARNING RENDSZERBEN

TOMPA TAMÁS
Miskolci Egyetem
Informatikai Intézet
Általános Informatikai Intézeti Tanszék
tompa@iit.uni-miskolc.hu

KOVÁCS SZILVESZTER
Miskolci Egyetem
Informatikai Intézet
Általános Informatikai Intézeti Tanszék
szkovacs@iit.uni-miskolc.hu

Absztrakt. A heurisztikusan gyorsított FRIQ-learning (HFRIQ-learning) a 'FIVE' fuzzy szabály-interpolációs modellen alapuló Q-tanuló módszer, amely alkalmas külső szakértői tudásbázis (mint szakértői heurisztika) injektálásra a rendszerbe. A beágyazott szakértői tudásbázis gyorsíthatja a tanulási folyamatot, de abban az esetben, ha ez az a priori tudásbázis helytelen információt tartalmaz, akkor az negatívan hathat a tanulási folyamat hatékonyságára. A HFRIQ-learning rendszerben a tudásbázist (Q-függvényt) egy állapot-akció-Q-érték formájú ritka (fuzzy szabály-interpolált) szabálybázis írja le, amely következtében a külső szakértői tudásbázis állapot-akció formájú fuzzy produkciós szabályok által adható meg. A cikk célja, hogy klasszikus megerősítéses tanulási mintapéldákon keresztül demonstrálja a HFRIQ-learning rendszerbe injektált szakértői tudásbázis hatását a tanulási folyamatra illetve annak bemutatása, hogy a rendszer hogyan valósítja meg a helytelen szakértői szabályok hangolását (optimalizálását).

Kulcsszavak: megerősítéses tanulás, heurisztikusan gyorsított megerősítéses tanulás, szakértői tudásbázis, Fuzzy szabály-interpoláció, Q-learning, FRIQ-learning

1. Bevezetés

A megerősítéses tanulás (Reinforcement Learning - RL) [15] olyan módszerek összesége melyek a probléma megoldását a környezet visszajelzései (megerősítései) alapján térképezik fel. Ezen algoritmusok próba-szerencse (trial and error) alapon törekednek arra, hogy az ágens teljesítményét javítsák (például maximalizálják a gyűjtött jutalmakat). A klasszikus RL algoritmusok, mint például a Q-learning [29], a SARSA [14], a Fuzzy Q-learning [3][7] illetve a mély Q-tanulás (Deep Q-learning) [6] tudásbázisa a tanulási folyamat elején teljesen üres, majd ezt a kezdetben üres tudásbázist töltik fel iterációról-iterációra. A tudásbázis (Q-függvény) reprezentáció módszerenként eltérő, Q-learning módszerek [29] esetében Q-tábla, fuzzy modellen alapuló módszerek esetében „ha-akkor" típusú szabályokat

tartalmazó fuzzy szabálybázis [7][10], fuzzy szabály-interpoláció alapú modell alkalmazása esetében pedig egy ritka szabálybázis [9][25][28].
A heurisztikusan gyorsított megerősítéses tanulás (Heuristically Accelerated Reinforcement Learning - HARL) [4] olyan módszerek összesége, amely alkalmas külső tudásbázis injektálására a rendszerbe. Ezen rendszerben a külső tudásbázis egy heurisztikus függvény által leírt, amely az adott állapotban preferált cselekvést (akciót) határozza meg az ágens számára. A HFRIQ-learning (Heuristically Accelerated Fuzzy Rule-Interpolation based Q-learning) [22] rendszerben az a priori szakértői tudásbázis egy ritka, állapot-akció formátumú, produkciós szabályokat tartalmazó fuzzy szabálybázis által leírt [18]. A megadott szakértői szabályok helyessége és azok száma (szakértői szabályrendszer minősége) hatással van a rendszer tanulási folyamatának hatékonyságára [17][19]. Egy szakértői szabály akkor tekinthető helyesnek, ha megfelelő akciót társít az adott állapothoz, amely következtében az ágens viselkedésére pozitívan hat. A HFRIQ-learning rendszer előnye, hogy alkalmas a helytelenül definiált (nem megfelelő akciót tartalmazó) szakértői szabályok keresésére és optimalizálására, a tanulási folyamat során alkalmazott gradiens módszer alapú hangolási eljárás illetve a szabálytávolság alapú szabálybázis csökkentési módszerek által. A cikk célja a helyes és helytelen szakértői tudásbázis hatásának bemutatása a HFRIQ-learning rendszerben klasszikus megerősítéses tanulási mintapéldák („Mountain car”, „Cart-Pole”) által.

2. Heurisztikusan gyorsított FRIQ-learning (HFRIQ-learning)

A heurisztikusan gyorsított FRIQ-learning (Heuristically Accelerated Fuzzy Rule-Interpolation based Q-learning - HFRIQ-learning) [22] a FRIQ-learning (Fuzzy Rule-Interpolation based Q-learning) [25] módszer kiterjesztése, amely által az alkalmas külső szakértői tudásbázis beágyazására [18] illetve hangolására [22]. A módszer a „FIVE” (Fuzzy Rule Interpolation based on Vague Environment) [11] fuzzy szabály-interpolációs eljárást alkalmazza a Q-függvény leírására, amely által az állapot-akció tér folytonos. A „FIVE” [11] egy alkalmazás-orientált FRI (Fuzzy Rule Interpolation) módszer, amely alacsony számítási igénye miatt [1][2] jól használható valós idejű alkalmazásokban illetve robotikai irányításokban. Továbbá, ezen alkalmazott FRI módszer által a rendszer komplexitása csökkenthető a ritka szabálybázis következtében illetve a rendszer abban az esetben is szolgáltat kimentet mikor a klasszikus fuzzy következtetési eljárások nem [12].
A HFRIQ-learning rendszer tudásbázisa egy ritka fuzzy szabálybázis által leírt, egy r_i ($i \in [1, m]$) szabály formája az m méretű R szabálybázisban a következő [25]:

$$r_i: \textbf{\textit{If}}\ s_1\ is\ S_1^i\ \textbf{\textit{And}}\ s_2\ is\ S_2^i\ \textbf{\textit{And}}\ \dots\ \textbf{\textit{And}}\ s_n\ is\ S_n^i\ \ \textbf{\textit{And}}\ a\ is\ A^i\ \textbf{\textit{Then}}\ \tilde{Q}(\boldsymbol{s}, a) = q^i \qquad (1)$$

ahol S_j^i az i-edik ($i \in [1, m]$) szabály j-edik ($j \in [1, n]$) állapot dimenziójának fuzzy halmaza az n-dimenziós $\boldsymbol{S}$ állapottérben, $\boldsymbol{s} \in \boldsymbol{S}$ az n-dimenziós állapot megfigyelés, s_j a j-edik dimenziója az $\boldsymbol{s}$ állapot megfigyelésnek, A^i az i-edik szabály egydimenziós akció univerzumának (U) fuzzy halmaza, $a \in U$ az akció, $\tilde{Q}(\boldsymbol{s}, a)$ a FIVE FRI [11] által becsült Q-függvény, q^i pedig az i-edik szabály konzekvense (Q-értéke).
Az R_{expert} szakértői tudásbázis formátuma hasonló az (1) formula által definiált fuzzy szabályokhoz, azzal az eltéréssel, hogy az $\hat{r}$ szakértői szabályok antecedense az állapot, konzekvense pedig az ebben az állapotban preferált akció [18]:

$$\hat{r}_i\text{: }\mathbf{If}\ s_1 \boldsymbol{is}\ \hat{S}_1^i\ \boldsymbol{And}\ s_2 \boldsymbol{is}\ \hat{S}_2^i\ \boldsymbol{And}\ \dots \boldsymbol{And}\ s_n \boldsymbol{is}\ \hat{S}_n^i\ \boldsymbol{Then}\ \text{a} = \hat{A}^i \tag{2}$$

ahol $\hat{r}_i$ az i-edik ($i \in [1, \hat{m}]$) szakértői szabály az R_{expert} szabálybázisban, $\hat{S}_n^i = \left[\hat{S}_1^i, \hat{S}_2^i, \dots \hat{S}_n^i\right]$ az i-edik szakértői szabály n-dimenziós állapot megfigyelése, $\hat{A}^i$ az ehhez az $\hat{S}_n^i$ állapot megfigyeléshez tartozó akció, i ($i \in [1, \hat{m}]$) pedig a szabály indexe az $\hat{m}$ méretű szakértői szabályrendszerben.
Annak következtében, hogy a szakértői szabályrendszer injektálható legyen a rendszerbe szükséges a formátumának átalakítása állapot-akció-Q-érték formátumra. Ekkor az átalakított szakértői szabályok antecedense az állapot-akció, konzekvense pedig egy becsült $\tilde{Q}_{init}$ érték lesz. A becsült kezdeti $\tilde{Q}_{init}$ érték a környezet által maximálisan adható megerősítés (g_{max}) ismeretében határozható meg [18]. A HFRIQ-learning rendszer tanulási folyamata ezen R_{expert} szakértői szabályrendszer és a 2^{n+1} darabszámú (n: állapotdimenziók száma), 0 konzekvens értékkel rendelkező $r_i^{\square}$ sarokponti szabályok összefésülésével létrejött fuzzy szabálybázissal kezdődik [18][25]. Abban az esetben ha ellentmondás alakul ki, azaz szakértői szabály sarokponti szabályra illeszkedik (de eltérő a konzekvensük), akkor az ellentmondás feloldása következtében ezen két szabály összevonásra kerül egyetlen szabállyá. A létrejött kezdeti szakértői szabályrendszer a tanulási folyamat során inkrementálisan növekszik a rendszer által létrehozott szabályokkal [26]. Új szabály akkor kerül beillesztésre a szabálybázisba, ha a Q-frissítés ($\Delta\tilde{Q}$) értéke nagyobb, mint egy ε_Q érték ($\Delta\tilde{Q} > \varepsilon_Q$) [26] és a legközelebbi szabálypont is távolinak tekinthető [21][22]. A szabályközelség meghatározásának az alapja a szabályok között definiált, dimenziónként számított távolságok [21][22]. Abban az esetben, ha a $\Delta\tilde{Q}$ érték kicsi ($\Delta\tilde{Q} < \varepsilon_Q$), akkor a teljes szabálybázis konzekvense kerül frissítésre a következő módon [26]:

$$\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \tilde{Q}^k(\boldsymbol{s}, a) + \Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) \tag{3}$$

$$\Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \alpha * \left(g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a') - \tilde{Q}^k(\boldsymbol{s}, a)\right) \tag{4}$$

ahol $\gamma \in [0,1]$ a leszámítolási tényező, $\alpha \in [0,1]$ a tanulási ráta, q_i^{k+1} az i-edik szabály konklúziója a $(k+1)$-edik iterációban, a pedig az $\boldsymbol{s}$-ben végrehajtott akció. Az új megfigyelt állapot $\boldsymbol{s}'$, $g(\boldsymbol{s}, a, \boldsymbol{s}')$ a megfigyelt jutalom az $\boldsymbol{s} \to \boldsymbol{s}'$ állapot átmenetre, $\tilde{Q}^k$ és $\tilde{Q}^{k+1}$ pedig a k-adik és a $(k+1)$-edik iteráció FIVE FRI módszer által becsült Q-értéke [25]:

$$\tilde{Q}(\boldsymbol{s}, a) = \begin{cases} \dfrac{q^i}{\left(\delta_v^i\right)^\lambda} \Big/ \left(\sum_{j=1}^{m} \dfrac{1}{\left(\delta_v^j\right)^\lambda}\right) & \text{ha } (\boldsymbol{s}, a) = (\boldsymbol{s}^i, a^i) \\ & \text{valamennyi } i - re, \\ \sum_{i=1}^{m} \left(\dfrac{q^i}{\left(\delta_v^i\right)^\lambda} \Big/ \left(\sum_{j=1}^{m} \dfrac{1}{\left(\delta_v^j\right)^\lambda}\right)\right) & \text{egyébként} \end{cases} \tag{5}$$

ahol q^i az i-edik ($i \in [1, m]$) szabály konklúziója, ($\mathbf{s}$, a) a megfigyelés, δ_v^i a skálázott távolság [11] az ($\mathbf{s}$, a) megfigyelés és az i-edik szabály ($\boldsymbol{s}^i, a^i$) antecedense között, λ a Shepard paraméter, m pedig a szabályok száma.
Ha a $\Delta\tilde{Q}$ értéke kicsi és van már létező szabály a megfigyelés közelében, akkor a gradiens módszer alapú hangolási eljárás a megfigyeléshez legközelebb elhelyezkedő szabálypont antecedensét és konzekvensét fogja hangolni [22]. A szabálypont új pozíciója az alábbi módon kerül meghatározásra [16][22]:

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(\boldsymbol{s}, a)}{\partial \boldsymbol{s}}\right) * \alpha \tag{6}$$

$$a_{k+1} = a_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(\boldsymbol{s}, a)}{\partial a}\right) * \alpha \tag{7}$$

$$q_{k+1} = q_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(\boldsymbol{s}, a)}{\partial q}\right) * \alpha \tag{8}$$

ahol a $\boldsymbol{s}_{k+1}, a_{k+1}, q_{k+1}$ a gradiens-módszer által meghatározott új állapot, akció és Q-értékek, $\boldsymbol{s}_k, a_k, q_k$ a régi állapot, akció és Q-értékek, α a gradiens-módszer tanulási rátája, $\frac{\partial \tilde{Q}(\boldsymbol{s},a)}{\partial \boldsymbol{s}}, \frac{\partial \tilde{Q}(\boldsymbol{s},a)}{\partial a}, \frac{\partial \tilde{Q}(\boldsymbol{s},a)}{\partial q}$ a Q-függvény állapot, akció és Q-érték szerinti parciális deriváltjai, a $TDerror$ értéke pedig a következő [16][22]:

$$TDerror = g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a') - \tilde{Q}^k(\boldsymbol{s}, a) \tag{9}$$

Az alkalmazott hangolási módszer következtében, a szabálypontok vándorlása miatt előfordulhat olyan eset, hogy több szabálypont is közel kerül egymáshoz. Ebben az esetben az egymáshoz közel kerülő és ez által hasonló információt leíró szabálypontok egyesítésre kerülnek egyetlen szabállyá [21][22], ami által a szabálybázis mérete a tanulási folyamat csökkenthető. A szabálytávolság alapú szabálybázis redukálási módszerről bővebb információ a [21][22][23] és [19] hivatkozásokban található. További, a tanulási folyamat után opcionálisan alkalmazható szabálybázis csökkentési módszereket a [20][24][27] hivatkozások mutatnak be.
A HFRIQ-learning tanulási folyamata akkor ér véget, ha nem kerül új szabály hozzáadásra az inkrementális szabálybázisba, a Q-frissítés értéke elenyészően kicsi, a létező szabálypontok pozíciója nem változik, és nem kerülnek szabályok összevonásra.

3. Alkalmazáspéldák

Ezen fejezet a bemutatott HFRIQ-learning módszer működésének hatékonyságát mutatja be egy egyetlen állapot- és egyetlen akcióváltozóval rendelkező mintapéldán és a klasszikus „Mountain car” illetve „Cart-Pole” megerősítéses tanulási alkalmazáspéldákon keresztül.

3.1. Egy állapot-akció változós mintapélda

A mintapélda nem egy klasszikus megerősítéses tanulási alkalmazáspélda, hanem a Q-függvény egyszerűbb vizualizációjának érdekében, egyetlen állapot és egyetlen akciódimenzióval rendelkező feladat. Az s_1 állapotváltozó értéktartománya -10 és +10 közé esik ($s_1 \in [-10,10]$), az a akcióváltozó pedig -2 és +2 között vehet fel értékeket ($a \in [-2,2]$), a tanulási módszer α paramétere 0.5 értékű, a γ diszkontálási tényező értéke 0.4, az ε-mohó akcióválasztás ε értéke pedig 0.5. A szakértő által meghatározott tudásbázis egyetlen állapot-akció formátumú szabályt tartalmaz, amely a 0 állapotpontban 0 értékű akciót határoz meg, és az erre megadott megerősítés értéke $g_{max} = 1$. A környezet +1 jutalmat ad, ha az ágens

állapotváltozójának értéke -1 és +1 közötti, ellenkező esetben a jutalom értéke 0. A jutalomfüggvény a következő:

Jutalomfüggvény: 1D_mintapélda

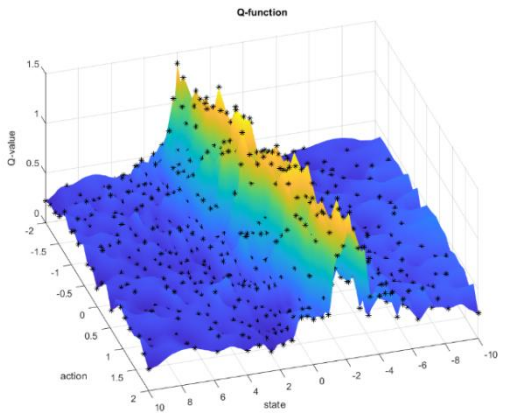
Bemenet: s_1 állapot
Kimenet: r megerősítés

```
if (s1 <= 1 and s1 >= -1)
    r = 1
else
    r = 0
end

return r
```

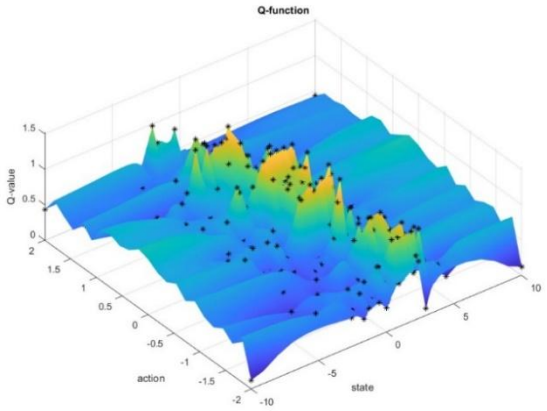
1. jutalomfüggvény:
Az egy állapot-akció változós mintapélda jutalomfüggvényének pszeudokódja

Az így kapott HFRIQ-learning futási eredmények összehasonlításának alapja a FRIQ-learning azon verziójának futása során kapott szabálybázis méret (szabályszám) illetve Q-függvény forma, ahol az eredeti állapot-akció tér rácsháló elhagyásra került és a szabályok között megengedett minimális szabálytávolság volt csak figyelembe véve. Az első futási esetben tehát a FRIQ-learning eredeti verziójában került futtatásra a mintapélda 10000 iteráción keresztül, sem a szakértői tudásbázis, sem a javasolt gradiens módszer alapú hangolási eljárás, sem pedig a szabálytávolság alapú szabálybázis redukálási módszer nem került alkalmazásra. Ebben az esetben a kapott szabálybázis 530 szabályt tartalmazott, amelyek által leírt Q-függvényt a következő 1. ábra látható „referencia” Q-függvény szemléltet:



1. ábra Az összehasonlítás alapjául szolgáló „referencia” Q-függvény

A második futási esetben egyetlen darab szakértői szabály lett a rendszerbe illesztve, illetve alkalmazásra került a javasolt HFRIQ-learning módszer. A szakértői szabály a 0 állapot pozícióban preferált 0 értékű akciót definiálja. A szabálybázis építési folyamat során a szabályok között megengedett minimális szabálytávolság az univerzumok méretének a 200-ad része ($\mathrm{dR}_{\boldsymbol{S}} = \mathrm{dR}_U = 200$). A szabálytávolság alapú szabálybázis redukálás dR paramétereinek értéke: $\mathrm{dR}_{\boldsymbol{S}} = 45$, $\mathrm{dR}_U = 45$ és $\mathrm{dR}_q = 100$. Ebben a futási esetben a szabálybázis 327 szabályt tartalmazott, amelyek által leírt Q-függvény a 2. ábrán látható:



2. ábra: A fejlesztett HFRIQ-learning rendszer alkalmazása esetében kapott Q-függvény

A 2. ábrán látható, hogy az így kapott Q-függvényt kevesebb tartópont (szabály) írja le a szabálytávolság alapú szabályösszevonás következtében, illetve a függvény formája a gradiens módszer–alapú szabálypont optimalizálás miatt kisimult a 1. ábra látható referenciaként szolgáló Q-függvényhez képest. Az 1. táblázat az egyes futási esetekben kapott szabálybázis méreteket (szabályszámokat) foglalja össze:

1. táblázat: Az egyes futási esetekben kapott szabálybázis méretek

#	Futási eset	Szabálybázis méret
1	szakértői tudásbázis nélkül illetve a gradiens módszer alapú hangolási eljárás és a szabálytávolság alapú szabálybázis redukálás nélkül	530 szabály
2	szakértői tudásbázissal és a gradiens módszer alapú hangolási eljárás és a szabálytávolság alapú szabálybázis redukálási módszerek alkalmazásával	327 szabály

A 2. táblázatban a tanulási fázis előtt, a szakértői által eredetileg definiált szabály illetve a tanulási fázis (a javasolt hangolási módszerek alkalmazása) után előállt szakértői szabály található:

2. táblázat:
A szakértő által definiált eredeti, illetve a hangolási folyamat után előállt szabály

#	Szakértői szabály	állapot	akció	Q-érték
1	eredetileg (tanulási fázis előtt) megadott	0	0	0.1
2	hangolt (tanulási fázis után)	0.06	0	0.59

Mivel a példában a szakértői szabály helyesen definiált volt, ezért a hangolás során a szabály állapot-akció pontja (azaz a szakértői szabály maga) csak kismértékben (akció értéke egyáltalán nem) változott (lásd 2. táblázat 1., 2. sora). A szabálypont hangolása után előállt pozitív Q-érték alapján a szabály helyes, a Q-érték változása csak a szabály hasznosságára vonatkozó becslést pontosítja.

3.2. „Mountain car” mintapélda

Ezen alfejezet egy klasszikus megerősítéses tanulási mintapéldán keresztül mutatja be a javasolt rendszer működését, hatékonyságát.
A választatott mintapélda a „Mountain Car” elnevezésű. Az alkalmazáspéldában az ágens egy autó, környezete pedig egy meredek völgy. Az autó a meredek völgy közepén helyezkedik el a tanulási folyamat indulásakor. Az ágens célja, hogy kijusson a meredek völgy közepéből a völgy tetején található dombra. A feladat akkor tekinthető megoldottnak, ha az autó, azaz az ágens valamennyi meghatározott lépés alatt (jelen esetben 2000 lépés alatt) kijut a völgyből és eléri a domb tetején lévő csillagot. A környezettől akkor érkezik nagy megerősítés (r) ha ez a feladat sikerül, azaz az ágens pozíciója eléri a csillag pozícióját, ellenkező esetben pedig büntetést ad a rendszer. A „Mountain Car” nevezetű megerősítéses tanulási probléma állapottere 2 változós (s_1, s_2), az akció tér pedig 1 változóval (a) rendelkezik, melyek a következők:

- autó aktuális pozícióját: s_1 ($s_1 \in [-1.5,0.5]$)
- autó aktuális sebessége: s_2 ($s_2 \in [-0.07,0.07]$)
- az autó elmozdulása: a (jobbra, balra vagy nincs elmozdulás, $a = [-1,0,1]$)

Ebben az esetben a tanulás paramétereinek értéke $\alpha = 0.5$, $\gamma = 0.99$, illetve egy epizód 2000 iterációból áll.

Az összehasonlítás alapjául három futási esetet hoztam létre, amelyek a következők:

- I. Eredeti FRIQ-learning verzió [25] (melyben az állapot-akció tér rácsháló alkalmazásra kerül, amely által az új szabályok állapot-akció pontja meghatározott).
- II. Az előző 1. eset kiegészítve szakértői tudásbázissal a szakértői tudásbázis leírási forma és az előzetes Q-érték meghatározási módszerek alkalmazásával.
- III. Az előző második eset, szakértői tudásbázis injektálásával és a javasolt HFRIQ-learning módszer alkalmazásával (az állapot-akció rácsháló elhagyásával, illetve javasolt gradiens módszer alapú hangolási eljárás és a szabálytávolság alapú szabálybázis redukálási módszerek alkalmazásával). Ez az futási eset négy további esetet takar az injektált szakértői tudásbázis milyenségétől függően:
    a) helyesen definiált szakértői szabályrendszer,
    b) részben helyesen definiált szakértői szabályrendszer (a helyesen megadott szakértői szabályoknak csak egy része),
    c) részben helytelenül definiált szakértői szabályrendszer,
    d) „véletlenszerűen” generált szakértői szabályrendszer.

Az összehasonlítás alapja, a rendszer tanulási hatékonyságára jellemző változók, azaz a konvergencia sebesség (a betanuláshoz szükséges epizódok száma), a megoldás megvalósításához szükséges lépések (step) száma (mennyi lépés alatt jut ki a völgyből az ágens) és a tudásbázis, azaz a fuzzy szabálybázis mérete.

Az első futási (I.) esetben a FRIQ-learning 29 epizód alatt konvergál és a szabálybázis 110 darab fuzzy szabályt tartalmaz (szabálybázis redukálás nélkül). Ezzel a tudásbázissal az ágens („kisautó”) 472 lépés (step) alatt jutott ki a völgyből és érte el a jobb oldali dombtetőn található sárga csillagot.

A második esetben (II.) kapott futási eredményeket a következő táblázat tartalmazza, melyről bővebb információk a [17][18] hivatkozásokban:

3. táblázat A II. futási esetben kapott eredmények

#	**Szakértői heurisztika típusa**	**Átlagos konvergencia sebesség**	**Átlagos szabályszám**
0.	üres (heurisztika nélkül)	28.3	91.7
1.	helyesen megadott	10	124.3
2.	helyesen megadottnak egy része	14.4	114.3
3.	részben helytelenül megadott	11.7	120.1
4.	véletlenszerűen generált	26.6	124.4

A harmadik futási esetben (III.) négy további esetet hoztam létre, azaz helyes (a.), részben helyes (b.), részben helytelen (c.), majd véletlenszerű szakértői szabályrendszerrel (d.) futott a HFRIQ-learning [16][21][22][23].

A futás során alkalmazott paraméterek értékei az alábbiak:

- gradiens módszer $\alpha = 0.01$
- új szabály felvételénél a szabályok közötti minimális szabálytávolságot meghatározó dR paraméterek értékei:
  - $\mathrm{dR}_{\boldsymbol{S}} = \mathrm{dR}_U = 40$
- a tanulási folyamat során alkalmazott szabálytávolság alapú szabálybázis redukálási módszer dR paramétereinek értékei:
  - $\mathrm{dR}_{\boldsymbol{S}} = 15, \mathrm{dR}_U = 15, \mathrm{dR}_q = 100$

Az egyes futási esetekben kapott eredményeket a 4. táblázat foglalja össze, ahol a III.a.-III.d.-vel jelölt futási esetek a HFRIQ-learning alkalmazása során kapott eredmények:

4. táblázat: Az egyes futási esetekben kapott eredmények

Futási eset	**Konvergencia sebesség (epizódok száma)**	**Megoldáshoz szükséges lépések száma**	**Szabálybázis méret (szabályok száma)**
I.	29	472	110
II.a.	10	NA	124.3
II.b.	10.4	NA	114.3
II.c.	11.7	NA	120.1
II.d.	26.6	NA	124.4
III.a.	1	1199	79
III.b.	9	1997	81
III.c.	20	1554	88
III.d.	37	1178	86

A táblázatban lévő futási eredmények alapján elmondható, hogy a tanulási folyamat konvergencia sebességét (és részben a végső tudásbázis méretét is) jelentős mértékben befolyásolja a szakértő által állapot-akció formátumban megadott tudásbázis helyessége. Ennek oka, hogy a helytelenül (vagy részben helytelenül) megadott előzetes tudásbázis esetében a szakértői szabályokat javítani (hangolni) kell. A szintén a szakértő által meghatározott dR paraméterek értékei alapján lettek kiszámítva a szabályfelvétel és a szabályösszevonási módszerek távolságküszöbeinek értékei, melyek a két szabály közötti minimális

szabálytávolságot határozzák meg, illetve a szabályösszevonás folyamatát irányítják. Ezen paraméter értékek függvényében mindig a megfigyeléshez legközelebbi szabálypont kerül hangolása. Ezért több iterációra (ennek következtében több időre) van szükség ahhoz, hogy a helytelenül megadott szakértői szabályok hangolására is sor kerüljön.

Abban az esetben (III.a.) amikor a helyesen definiált szakértői szabályrendszerrel futott a szimuláció, akkor a rendszer jelentősen gyorsabban konvergált (egyetlen epizód alatt), mint a FRIQ-learning eredeti verziójában és a szabályszám is csökkent, 110-ről 79-re. Ennek oka, hogy a megadott szakértői tudásbázison nem kellett hangolnia a rendszernek, a szabályszámot pedig a tanulási folyamat közben alkalmazott szabálytávolság alapú szabálybázis redukciós módszer csökkentette.

Abban az esetben mikor a részben helyes szakértői szabályokkal futott a tanulási folyamat (III.b.), akkor a rendszer 9 epizód alatt konvergált 81 darab szabállyal, tehát valamennyivel több epizódra volt szükség, mint az előző futási esetben.
Mikor a részben helytelen szakértői szabályokkal futott a szimuláció (III.c), azaz több elrontott (helytelen) szabályt is tartalmazott a szakértői tudásbázis, akkor 20 epizódra volt szükség ahhoz, hogy a javasolt módszer kijavítsa a helytelenül megadott szakértői szabályokat.
Mikor a tejesen helytelen előzetes tudásbázis lett injektálva a tanulási folyamatba (III.d.), akkor is konvergált a rendszer de a 37 epizódra (epizódonként 2000 iterációra) volt szükség ahhoz, hogy a helytelen szakértői szabályok hangolása megvalósuljon.

A következő táblázatok a „véletlenszerűen” generált (helytelen) szakértői szabályrendszer szabályait tartalmazzák a tanulási fázis előtt (5. táblázat) és a tanulási fázis után (6. táblázat), azaz a javasolt szabálybázis hangolási (és redukálási) módszerek alkalmazását követően:

5. táblázat:
A „véletlenszerűen” generált (helytelen) szakértői szabályrendszer szabályai a tanulási fázis előtt

R#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
s1	-0.475	-0.5	-0.475	-0.475	-0.27	-0.27	-0.27	-0.475	-0.475
s2	0	0	-0.014	0.014	0	-0.014	0	-0.042	0
a	1	-1	-1	0	-1	0	-1	1	1

R#	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
s1	-0.475	-0.065	0.14	-0.27	-0.885	0.885	-0.065	-1.09
s2	0	0	-0.014	-0.042	0.042	0.042	0.042	0.042
a	-1	0	1	-1	-1	1	0	-1

6. táblázat:
A „véletlenszerűen" generált (helytelen) szakértői szabályrendszer szabályai a tanulási (hangolási) fázis után

R#	1	2	3	4	5	6	7	8	9
s1	x	x	x	x	x	x	-0.52	x	-0.39
s2							-0.04		0.05
a							-1		1

R#	10	11	12	13	14	15	16	17
s1	x	-0.21	-0.47	-0.31	x	0.885	x	-0.81
s2		-0.03	-0.016	-0.03		0.042		0.03
a		0	1	-1		1		-1

A 6. táblázat alapján látható, hogy a tanulási folyat során a 17 darab helytelen szakértői szabályból csupán 7 darab szakértői szabályt (7, 9, 11, 12, 13, 15, 17 sorszámú szabályok) hagyott meg a módszer, mely 7 szakértői szabály közül 6 darab állapot-akció pontját jelentősen elhangolta (optimalizálta), a 15. szakértői szabályt viszont érintetlenül hagyta. Ezek alapján megállapítható, hogy a kezdeti szakértői szabályrendszerben csak 15. sorszámú az igazoltan helyesen definiált szakértői szabály. Az „x" jelölésű szakértői szabályok a szabálybázis redukálás, azaz az a szabályösszevonások során törlésre kerültek annak következtében, hogy összeolvadtak más szakértői, vagy újonnan beszúrt szabályokkal. A szabályok száma olyan módon csökkent, hogy összeolvadtak más szabályokkal a szabálybázis hangolás és a szabálybázis redukálás során, így a szabálybázisból elhagyott (törölt) szabályok redundáns szabályoknak tekinthetők.
A kapott futási eredmények alapján megállapítható, hogy a javasolt HFRIQ-learning rendszer (és így a javasolt szabálybázis hangolási és redukálási módszerek) alkalmazása által a rendszerbe injektált szakértői tudásbázis hangolható (korrigálható) azokban az esetekben mikor az helytelen információkat tartalmaz.

3.3. „Cart-Pole" mintapélda

A javasolt HFRIQ-learning rendszer működése és hatékonysága egy újabb klasszikus megerősítéses tanulási mintapéldán, a „Cart-Pole" nevezetű szimulációs példán keresztül kerül vizsgálatra.
Ebben az esetben az ágens egy autó melynek célja, hogy a közepén elhelyezkedő rudat megtanulja függőleges pozícióban tartani. A probléma 4 állapotleíróval és 1 akcióváltozóval rendelkezik, melyek a következők:

- autó aktuális vízszintes pozíciója: s_1
- autó aktuális sebessége: s_2
- inga aktuális pozíciója (szög): s_3
- inga szögsebessége: s_4
- autó elmozdulása adott erővel: a (jobbra, balra vagy nincs elmozdulás)

A rendszer a futási paramétereinek értékei: $\alpha = 0.3,\ \gamma = 0.99$. A vizsgált futási esetek az injektált szakértői tudásbázis milyenségétől függően a futási esetek a

következők:

1. szakértői tudásbázis nélkül (FRIQ-learning)
2. helyes szakértői szabályrendszer
3. részben helyes szakértői szabályrendszer
4. teljesen helytelen szakértői szabályrendszer

Az összehasonlítás alapja a szakértői tudásbázis nélküli 1. futási eset konvergencia sebessége és szabálybázis mérete, amikor az eredeti FRIQ-learning rendszer [25] 58 epizóddal és 182 darab fuzzy szabállyal konvergált.
Azon esetekben kapott futási eredményeket mikor a FRIQ-learning rendszer beágyazott szakértői tudásbázissal, de a gradiens módszer alapú hangolási eljárás és a szabálytávolság alapú szabálybázis redukálási módszerek alkalmazása nélkül került futtatásra, a következő táblázat foglalja össze:

7. táblázat Futási eredmények a FRIQ-learning rendszer esetében de beágyazott szakértői tudásbázissal

#	**Szakértői heurisztika típusa**	**Konvergencia (epizód szám)**	**Szabálybázis méret**
1.	üres	58	182
2.	helyes	5	46
3.	részben helyes	65	263
4.	teljesen helytelen	nem konvergál	>400

A HFRIQ-learning esetében a szabálybázis építési folyamat során a szabályok között megengedett minimális szabálytávolság az univerzumok méretének a 10-ed része ($\mathrm{dR}_{\boldsymbol{S}} = \mathrm{dR}_U = 10$), a szabálytávolság alapú szabálybázis redukálás dR paramétereinek értéke $\mathrm{dR}_{\boldsymbol{S}} = 4$, $\mathrm{dR}_U = 5$, $\mathrm{dR}_q = 10$, a gradiens módszer alapú hangolási eljárás α tanulási ráta paraméterének értéke pedig $\alpha = 0.01$. Az ebben az esetben kapott futási eredményeket a következő 8. táblázat foglalja össze:

8. táblázat: Az egyes futási esetekben kapott eredmények

Futási eset	**Szakértői heurisztika típusa**	**Konvergencia sebesség (epizódok száma)**	**Szabálybázis méret (szabályok száma)**
1.	üres	58	182
2.	helyes	2	74
3.	részben helyes	26	140
4.	teljesen helytelen	92	95

Ezen alkalmazáspélda esetében kapott futási eredmények alapján is elmondható, hogy a szakértő által megadott tudásbázis helyessége jelentős mértékben befolyásolja a HFRIQ-learning rendszer konvergencia sebességét és végső szabálybázisának méretét. Abban az esetben (2. eset) mikor helyes szakértői szabályrendszer került injektálásra akkor a rendszer 2 epizód alatt konvergált, aminek oka, hogy a rendszernek nem kellett hangolnia a megadott szakértői tudásbázison. Azokban az esetekben (3. és 4. eset) amikor helytelen szakértői szabályrendszer került injektálásra a rendszerbe akkor a javasolt hangolási módszer alkalmazása következtében a rendszernek több iterációra volt szükség a szabályok hangolásához.

A 7. táblázatban bemutatott eredmények szerint, a javasolt hangolási eljárás nélkül,

ugyanezen szakértői tudásbázis esetén 350 epizód után (400 darab szabállyal) sem konvergált a rendszer, nem találta meg a megoldást leíró tudásbázist. Ezt a hibát a javasolt szabálybázis hangolási (és a szabálytávolság alapú szabálybázis redukálási) módszer kiküszöbölte.
A következő táblázatok a „véletlenszerűen" generált (helytelen) szakértői szabályrendszer szabályait tartalmazzák a tanulási fázis előtt (9. táblázat) és a tanulási fázis után, azaz a javasolt HFRIQ-learning rendszer alkalmazását követően (10. táblázat):

9. táblázat: A teljesen helytelen szakértői szabályrendszer szabályai a tanulási fázis előtt

R#	**1**	**2**	**3**	**4**	**5**	**6**	**7**
$\mathbf{s_1}$	1	1	1	1	-1	-1	1
$\mathbf{s_2}$	0	0	0	0	0	1	0
$\mathbf{s_3}$	0	-0.0524	0	-0.0524	-0.2094	-0.2094	0.2094
$\mathbf{s_4}$	1	-1	-1	1	-1	-1	1
a	-1	1	-1	-1	0.8	0.4	1

10. táblázat: A teljesen helytelen szakértői szabályrendszer szabályai a tanulási (hangolási) fázis után

R#	**1**	**2**	**3**	**4**	**5**	**6**	**7**
$\mathbf{s_1}$	0.9340	1.0115	1.0095	0.9630	-1	-1	
$\mathbf{s_2}$	-1.3334	0.5245	0.1166	-0.3350	0	1	
$\mathbf{s_3}$	0.0169	-0.0173	-0.0156	0.0819	-0.2094	-0.2094	x
$\mathbf{s_4}$	1.520	-0.7900	-0.1966	0.6808	-1	-1	
a	-0.9475	0.9692	-0.9108	0.975	0.8	0.4	

A 10. táblázatban látható, hogy a 7 darab helytelenül megadott szakértői szabály közül az első 4 darab (1.-4.) szabályt hangolta a rendszer, az 5. és 6. szabályokat változatlanul hagyta, a 7. szabály pedig összevonta más szabállyal (törölte). A tanulási folyamat ebben az esetben 92 epizódot vett igénybe és a végső szabálybázis 95 szabályt tartalmazott.

4. Köszönetnyilvánítás

„A Kulturális és Innovációs Minisztérium ÚNKP-23-4-I-ME/5 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült."

5. Összefoglalás

Bemutatásra került a HFRIQ-learning (Heuristically Accelerated Fuzzy Rule-Interpolation based Q-learning) rendszer, amely egy olyan fuzzy szabály-interpolációs módszeren alapuló Q-learning módszer, ami alkalmas külső kezdeti szakértői tudásbázis injektálására és hangolására (optimalizálására) majd annak visszaolvasására a tanulási folyamat végeztével.
A mintapéldák által kapott futási eredmények alapján elmondható, hogy a bemutatott HFRIQ-learning rendszer tanulási fázisának konvergencia sebessége javulhat, de csak olyan esetekben mikor a szakértői heurisztika helyes. Ellenkező esetben, mikor helytelen szakértői szabályok is injektálásra kerülnek, a tanulási módszer továbbra is konvergál, de a szabályok hangolásához (optimalizálásához) több epizódra (és így több iterációra) van szükség.
További kutatási terv egy olyan szakértői tudásbázis validációs módszer kidolgozása, amely alkalmas lehet a szakértő által megadott kezdeti, állapot-akció típusú fuzzy szabályok és a tanulási (hangolási) folyamat végeztével előállt optimalizált szakértői szabályok összehasonlítására, információt adhat arról, hogy a kezdeti szakértői szabályrendszer milyen mértékben voltak helyesek.

Irodalomjegyzék

[1] Bartók, Roland, and József Vásárhelyi. "Design of a FPGA accelerator for the FIVE fuzzy interpolation method." International Journal of Computer Applications in Technology 68.4 (2022): 321-331. https://doi.org/10.1504/ijcat.2022.125185

[2] Bartók, Roland, and József Vásárhelyi. "Examining Cache Handling of the FIVE Method on Multicore Systems." 2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2019. https://doi.org/10.1109/sami.2019.8782721

[3] Berenji, Hamid R. "Fuzzy Q-learning for generalization of reinforcement learning." Proceedings of IEEE 5th International Fuzzy Systems. Vol. 3. IEEE, 1996.

[4] Bianchi, Reinaldo AC, Carlos HC Ribeiro, and Anna Helena Reali Costa. "Heuristically Accelerated Reinforcement Learning: Theoretical and Experimental Results." ECAI. 2012. https://doi.org/10.1109/fuzzy.1996.553542

[5] Chen, G. and Yhou, J.: Boundary Element Methods. Academic Press Limitid, 24-28 Oval Road, London, NW1 7DX, 1992, ISBN 0-1-170840-X.

[6] Fan, Jianqing, et al. "A theoretical analysis of deep Q-learning." Learning for Dynamics and Control. PMLR, 2020. https://doi.org/10.48550/arXiv.1901.00137

[7] Glorennec, P. Y., and L. Jouffe. "Fuzzy Q-Learning Proc. of FUZZ-IEEE'97." (1997). https://doi.org/10.1109/fuzzy.1997.622790

[8] Gurtin, M. E.: The Linear Theory of Elasticity. In S. Flügge (ed.), Handbuch der Physik, Festkörpermechanik, vol. 2, pp. 57-60, Springer Verlag, Berlin, Heidleberg, NewYork, 1st edn., 1972. https://doi.org/10.1126/science.125.3239.162.b

[9] Horiuchi, T., Fujino, A., Katai, O., & Sawaragi, T. (1996, September). Fuzzy interpolation-based Q-learning with continuous states and actions. In Proceedings of IEEE 5th International Fuzzy Systems (Vol. 1, pp. 594-600). IEEE. https://doi.org/10.1109/fuzzy.1996.551807

[10] Kim, Min-Soeng, Gun-Gi Hong, and Ju-Jang Lee. "Online fuzzy Q-learning with extended rule and interpolation technique." Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289). Vol. 2. IEEE, 1999. https://doi.org/10.1109/iros.1999.812771

[11] Kovács, Szilveszter. "Extending the fuzzy rule interpolation" FIVE" by fuzzy observation." Computational Intelligence, Theory and Applications. Springer, Berlin, Heidelberg, 2006. 485-497. https://doi.org/10.1007/3-540-34783-6_48

[12] Kovacs, Szilveszter. "Fuzzy Rule Interpolation in Practice." SCIS & ISIS SCIS & ISIS 2006. Japan Society for Fuzzy Theory and Intelligent Informatics, 2006. https://doi.org/10.14864/softscis.2006.0.256.0

[13] Paulino, G. H.: Novel Formulations of the Boundary Element Method for Fracture Mechanics and Error Estimation. Ph. D. Dissertation, Cornel University, Ithaca, NY, USA, 1995.

[14] Rummery, Gavin A., and Mahesan Niranjan. On-line Q-learning using connectionist systems. Vol. 37. Cambridge, UK: University of Cambridge, Department of Engineering, 1994.

[15] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018. https://doi.org/10.1017/s0263574799211174

[16] Tamás, Tompa, and Kovács Szilveszter. "Expert heuristic tuning design for the FRIQ-learning." Multidiszciplináris Tudományok 10.4 (2020): 119-125. https://doi.org/10.35925/j.multi.2020.4.15

[17] Tompa, T., Kovács, S., Vincze, D., & Niitsuma, M. (2021, January). Demonstration of expert knowledge injection in Fuzzy Rule Interpolation based Q-learning. In 2021 IEEE/SICE International Symposium on System Integration (SII) (pp. 843-844). IEEE. https://doi.org/10.1109/ieeeconf49454.2021.9382734

[18] Tompa, Tamás, and Szilveszter Kovács. "Applying Expert Heuristic as an a Priori Knowledge for FRIQ-Learning." Acta Polytechnica Hungarica 17.4 (2020). https://doi.org/10.12700/aph.17.4.2020.4.2

[19] Tompa, Tamás, and Szilveszter Kovács. "Benchmark example for the Heuristically accelerated FRIQ-learning." 2023 24th International Carpathian Control Conference (ICCC). IEEE, 2023. https://doi.org/10.1109/iccc57093.2023.10178919

[20] Tompa, Tamás, and Szilveszter Kovács. "Clustering-based fuzzy knowledgebase reduction in the FRIQ-learning." 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2017. https://doi.org/10.1109/sami.2017.7880302

[21] Tompa, Tamás, and Szilveszter Kovács. "Determining the minimally allowed rule-distance for the incremental rule-base contruction phase of the FRIQ-learning." 2018 19th International Carpathian Control Conference (ICCC). IEEE, 2018.

[22] Tompa, Tamás, and Szilveszter Kovács. "Heuristically accelerated FRIQ-learning." 20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2022). IEEE, 2022. https://doi.org/10.1109/carpathiancc.2018.8399677

[23] Tompa, Tamás, and Szilveszter Kovács. "Tudásbázis redukálás a heurisztikusan gyorsított FRIQ-learning rendszerben." Production Systems and Information Engineering 11.2 (2023): 1-12. https://doi.org/10.32968/psaie.2022.4.4

[24] Vincze, Dávid, Alex Tóth, and Mihoko Niitsuma. "Antecedent redundancy exploitation in fuzzy rule interpolation-based reinforcement learning." 2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). https://doi.org/10.1109/aim43001.2020.9158875

[25] Vincze, Dávid, and Szilveszter Kovács. "Fuzzy rule interpolation-based Q-learning." 2009 5th International Symposium on Applied Computational Intelligence and Informatics. IEEE, 2009. https://doi.org/10.1109/saci.2009.5136311

[26] Vincze, Dávid, and Szilveszter Kovács. "Incremental rule base creation with fuzzy rule interpolation-based Q-learning." Computational Intelligence in Engineering. Springer, Berlin, Heidelberg, 2010. 191-203. https://doi.org/10.1007/978-3-642-15220-7_16

[27] Vincze, Dávid, and Szilveszter Kovács. "Rule-base reduction in Fuzzy Rule Interpolation-based Q-learning." Recent Innovations in Mechatronics 2.1-2. (2015): 1-6. https://doi.org/10.17667/riim.2015.1-2/10.

[28] Vincze, Dávid. "Fuzzy rule interpolation and reinforcement learning." 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics

(SAMI). IEEE, 2017. https://doi.org/10.1109/sami.2017.7880298

[29] Watkins, Christopher JCH, and Peter Dayan. "Q-learning." Machine learning 8.3 (1992): 279-292. https://doi.org/10.1023/a:1022676722315