Multidiszciplináris tudományok, 9. kötet. (2019) 4 sz. pp. 356-368 **https://doi.org/10.35925/j.multi.2019.4.35**

SZAKÉRTŐI HEURISZTIKA ALKALMAZÁSA A FRIQ-LEARNING MEGERŐSÍTÉSES TANULÁSI MÓDSZERBEN

Tompa Tamás
tanársegéd, Miskolci Egyetem, Informatikai Intézet, Általános Informatikai Intézeti Tanszék
Cím: 3515 Miskolc, Miskolc-Egyetemváros, e-mail: tompa@iit.uni-miskolc.hu

Kovács Szilveszter
docens, Miskolci Egyetem, Informatikai Intézet, Általános Informatikai Intézeti Tanszék
Cím: 3515 Miskolc, Miskolc-Egyetemváros, e-mail: szkovacs@iit.uni-miskolc.hu

Absztrakt

Jelen cikk szakértői tudásbázis, mint előzetes, a priori heurisztika alkalmazási lehetőségét és annak hatását mutatja be a FRIQ-learning megerősítéses tanulási módszerben. A megerősítéses tanulási módszerek többsége, mint ahogyan a FRIQ-learning rendszer is üres tudásbázissal indítja a tanulási folyamatot, majd egy megfelelően meghatározott jutalomfüggvény alapján inkrementálisan bővíti azt. A cikk bemutatja a FRIQ-learning algoritmus továbbfejlesztett verzióját, amely esetében a rendszer nem üres tudásbázissal indítja a tanulási fázist, hanem egy szakértő által megadott, előzetes tudásbázissal. A bemutatott módszer segítségével az előzetes szakértői heurisztika beágyazható az FRIQ-learning módszerbe. Továbbá a cikk a népszerű „mountain car" mintapéldán keresztül szemlélteti a szakértői tudásbázis beágyazásának módját és hatását a rendszerre.

Kulcsszavak: *megerősítéses tanulás, Q-learning, fuzzy szabály-interpoláció, fuzzy Q-learning, szakértői tudásbázis*

Abstract

The paper introduces expert knowledge as a priori heuristic application and its effect on the FRIQ-learning methodology. In general, the reinforcement learning methods, like the FRIQ-learning system, start with an empty knowledge base then the system builds the final knowledge base incrementally by the properly defined reward function. The main goal of the paper is to introduce the new developed version of the FRIQ-learning. In this case, the system starts the learning phase with not an empty knowledgebase but with an expert-defined, a priori knowledgebase. The introduced methodology is suitable for adapt expert knowledge in the FRIQ-learning system. Furthermore, the expert knowledge adaptation and its effect on the system are also discussed in the paper through the „mountain car" application example.

Keywords: *reinforcement learning, Q-learning, fuzzy rule interpolation, fuzzy Q-learning, expert knowledgebase*

1. Bevezetés

A növekvő számítási kapacitás nyújtotta lehetőségek következtében a mesterséges intelligencia egyre inkább megjelenik hétköznapi eszközeinkben. A gépi tanulás témaköre így egyre jobban növekvő jelentőséggel bíró tudomány terület, melynek egyik népszerű tématerülete a megerősítéses tanulás (Reinforcement Learning – RL) [8]. Ezek a gépi tanulási módszerek jól használhatóak olyan rendsze-

rekben ahol nem áll rendelkezésre, vagy esetleg csak részlegesen a rendszert működtető tudásbázis és így nem ismert a működés tényleges folyamata. Ebben az esetben tudáskinyerésre, tudásbázis létrehozására van szükség, melynek egyik megvalósítási módja lehet valamilyen megerősítéses tanuló módszer alkalmazása.

A megerősítéses tanulási eljárások működésének az alapötlete, hogy az ágens egy, az adott problémához megfelelően formált jutalomfüggvény segítségével, majd annak következtében a környezetből érkező, adott mértékű jutalmak és büntetések által, folyamatosan próbálkozva igyekszik feltárni az adott megoldást. Ezen módszerek esetében az elérendő cél jutalomfüggvény formájában van definiálva, amely helyes megválasztása kulcsfontosságú, ennek alapján fogja meghatározni a rendszer, hogy az ágens által végrehajtott cselekvések közül mely volt helyes és mely nem.

Számos megerősítéses tanulási algoritmus található a szakirodalomban, ezek közül a legelterjedtebb a Q-learning [16] és annak különböző változatai, például a Fuzzy Q-learning [1]. Ezen algoritmusok mindegyike a tanulási folyamatot üres tudásbázissal indítja el, kezdetben semmilyen tudással nem rendelkezik az adott probléma megoldására vonatkozóan. A tudásbázisát a jutalomfüggvény által építi minden egyes lépésben, iterációról-iterációra. Q-learning esetében a tudásbázis egy Q-tábla által van ábrázolva, melyben az adott állapotban végrehajtott cselekvésekhez tartozó jóságértékek (Q-értékek) vannak eltárolva, kezdetben minden érték nulla. Fuzzy alapú Q-learning esetében egy fuzzy szabálybázis írja le a rendszer működtető tudását oly módon, hogy a szabály antecedense (előzménye) az állapotok és a hozzátartozó akció, konzekvense (következménye) pedig a megfelelő Q-érték. Kezdetben a szabálybázis egy szabályt sem tartalmaz. A rendszert működtető végső tudásbázis létrehozásának folyamata hosszadalmas lehet, ezen eljárások lépésről-lépésre bővítik és hangolják azt. Azonban ha rendelkezésre áll a rendszert működtető tudás egy része és az valamilyen módon beágyazható a módszerbe, akkor a tanulási fázis lerövidíthető, függve az előzetesen megadott (részleges) tudásbázis helyességétől, méretétől.

Néhány hasonló módszer megtalálható a szakirodalomban, melyek előzetes tudásbázis alkalmazásával működtetik a rendszert. Egyik ilyen a 'Heuristically Accelerated Reinforcement Learning' (HARL) [3] témaköre, amely egy heurisztikus függvényt definiálva adja meg, hogy mely állapotokban mely cselekvés végrehajtása a legkedvezőbb. Egy másik megoldás az előzetes heurisztikát egy tudásreprezentációs nyelv, a 'GOAL' által írja le , amely az ágens számára „ha-akkor" típusú szabályok formájában határozza meg az akcióválasztási politikát [4][7] által javasolt módszer szerint Fuzzy Q-learning algoritmusban fuzzy szabályok formájában fogalmazható meg előzetes szakértői tudás.

Jelen cikk bemutat egy olyan eljárást, amely által a fuzzy szabály-interpoláció alapú Q-tanulás (Fuzzy Rule Interpolation-based Q-learning, FRIQ-learning) [11] [14] megerősítéses tanulási módszerbe előzetes (a priori), szakértő által megadott részleges tudásbázis, mint szakértői heurisztika beépíthető, a tanulási folyamat ezzel a szakértői heurisztikával indítható. A „mountain car" alkalmazáspéldán keresztül bemutatásra kerül továbbá a szakértői heurisztika megadásának jelenlegi módja és a szakértői tudásbázis rendszerre gyakorolt hatása.

2. Az FRIQ-learning megerősítéses tanulási módszer

A fuzzy szabály-interpoláció alapú Q-tanulás (FRIQ-learning) egy fuzzy szabály-interpolációs (Fuzzy Rule Interpolation, FRI) eljárást alkalmazó megerősítéses tanulási algoritmus. Az alkalmazott FIVE (Fuzzy Rule Interpolation based on Vague Environment) [5] szabály-interpoláció segítségével az eredeti diszkrét felbontású Q-learning algoritmus működését terjeszti ki folytonos állapot-akció (és Q-érték) térre. A módszer előnye a szakirodalomban is megtalálható fuzzy alapú Q-learning (FQ-learning

[2]) algoritmusokhoz képest, hogy FIVE FRI modell alkalmazása következtében a rendszer tudásbázisát egy ritka fuzzy szabálybázis írja le. Ennek következtében nincs szükség az összes lehetséges állapot-akció kombinációra fuzzy szabály felvételére, hanem elegendő csak a lényegi (kardinális) szabályokat létrehozni, jelentősen csökkentve ez által az adott probléma megoldását leíró tudásbázis méretét.

A rendszer működtető tudásbázisát leíró i-edik ($i \in [1,r]$, r a szabályszám) fuzzy szabály alakja a következő:

$$\textbf{If } s_1 \textbf{ is } S^i_1 \textbf{ And } s_2 \textbf{ is } S^i_2 \textbf{ And … And } s_n \textbf{ is } S^i_n \textbf{ And } a \textbf{ is } A^i \textbf{ Then } \tilde{Q}(\boldsymbol{s},a) = q^i \tag{1}$$

ahol, $\tilde{Q}(s,a)$ a közelített Q-függvény, q^i az i-edik szabály konzekvense. S^i_j ($j \in [1,n]$) a fuzzy halmaza az *i*–edik szabálynak a *j*–edik antecedens dimenzióban, az *n* dimenziós állapot térben $\boldsymbol{S}$ ($\boldsymbol{s} \in \boldsymbol{S}$) az *n* dimenziós állapot megfigyelés, s_j a *j*-edik dimenziója az állapot megfigyelés $\boldsymbol{s}$-nek, A^i a fuzzy halmaza az *i*-edik szabálynak az egydimenziós akciótérben *U*, *a* ($a \in U$) a végrehajtott akció. A rendszer állapot-akció tere *n+1* dimenziós, ahol *n* az állapot dimenziók száma, a további dimenzió pedig az akciótér miatt jelenik meg.

A FIVE FRI modellel közelített $\tilde{Q}(s,a)$-függvény frissítési formulája, ami az *i*-edik fuzzy szabály q_i-edik konzekvensét becsli a *(k+1)*-edik iterációban a következő:

$$q_i^{k+1} = \begin{cases} q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) & \text{if } (\boldsymbol{s},a) = (\boldsymbol{s}^i, a^i) \text{ valamenny i } i-re, \\ q_i^k + \Delta\tilde{Q}^{k+1}(\boldsymbol{s},a) \cdot \left(1/\delta^{\lambda}_{v,i}\right) \Big/ \left(\sum_{i=1}^{r} 1/\delta^{\lambda}_{v,i}\right) & \text{egyébként.} \end{cases} \tag{2}$$

$\Delta\tilde{Q}^{k+1}(\boldsymbol{s},a)$ a Q-függvény *(k+1)*-edik update értéke ($\boldsymbol{s}$,*a*) -ban a következő módon határozható meg:

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

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

Ahol, γ a leszámítolási tényező, az $\alpha \in [0,1]$ pedig a tanulási ráta. q_i^{k+1} az i-edik szabály singleton konklúziója a *(k+1)*-edik iterációban, *a* a végrehajtott akció $\boldsymbol{s}$-ben, $\boldsymbol{s}'$ az új állapot megfigyelés, $g(\boldsymbol{s},a,\boldsymbol{s}')$ a jutalom az $\boldsymbol{s} \to \boldsymbol{s}'$ állapotátmentre. A $\tilde{Q}^k$ és $\tilde{Q}^{k+1}$ értékek a *k*-adik és a *(k+1)*-edik iteráció becsült konklúziója, a FIVE FRI által (2).

A tanulási fázis kezdetben 2^{n+1} darabszámú fuzzy szabálybázissal indul, amit az inkrementális szabálybázis építési módszer [12] iterációról-iterációra bővít. Ezek az úgynevezett sarokponti vagy kezdeti szabályok konzekvens értékei rendre q_i=0 és az *n+1* dimenziós hiperkocka sarkaiban helyezkednek el. A továbbiakban a rendszer ezt a kezdeti szabálybázist bővíti vagy hangolja, attól függően, hogy szükséges-e új szabály felvétele a szabálybázisba vagy csak a meglévő tudásbázist kell frissíteni. Új szabály felvétele az ágens környezetéből érkező megerősítési információk és a Q-frissítési értékek

($\Delta\tilde{Q}$) alapján történik. Ha a Q-update értéke magasabb, mint egy előre meghatározott Q-update limit ($\Delta\tilde{Q} > \varepsilon_Q$) és a létező legközelebbi szabály is távol van az éppen beszúrandó szabály pozíciójához képest, akkor új szabály felvétele történik az adott lehetséges szabálypozícióba. A lehetséges szabálypozíciókat egy állapot-akció-tér rácsháló határozza meg ($s_{k+1} = s_k, \forall k > i, s_{i+1} = \frac{s_i + s_{i+2}}{2}$). Abban az esetben, ha a Q-update értéke kisebb, mint a meghatározott Q-frissítési limit ($\Delta\tilde{Q} < \varepsilon_Q$), akkor nem történik szabály beszúrás, hanem a teljes szabálybázis frissítése (hangolása) valósul meg oly módon, hogy a létező összes szabály konzekvense frissítésre kerül. Ez a lépés inkrementálisan valósul meg, minden egyes iterációban. Akkor ér véget a tanulási fázis és áll elő a végleges tudásbázis, ha a rendszer már nem illeszt be új szabályt a szabálybázisba és a Q-frissítési értékek relatívan kicsik maradnak.

Az inkrementális szabálybázis építési folyamat által létrehozott tudásbázis tartalmazhat olyan szabályokat, amelyeknek csak az építési fázisban volt szerepük vagy esetleg kiadódhatnak más szabályokból. Ezek a redundáns szabályok elhagyhatók a tanulási fázis végeztével előállt szabálybázisból, csökkentve ez által a végleges tudásbázis méretét. Az elhagyható, azaz a törölhető szabályok megállapítására 4 dekrementális tudásbázis redukálási módszerrel rendelkezik a FRIQ-learning rendszer [13], amelyek az inkrementális szabálybázis építési fázis után futtathatók opcionálisan. Az I. redukálási stratégia azokat a szabályokat törli a rendszerből elsődlegesen, amelyek a legkisebb Q-értékekkel rendelkeznek. Ezt minden egyes lépseben végrehajtja, majd vizsgája, hogy az így előállt szabályrendszer megoldja-e az adott problémát. A II. stratégia a nagyobb Q-értékekkel rendelkező szabályokat vizsgálja meg először. A III. stratégia szabálycsoportokat alakít ki Q-értékek alapján majd az így kialakult szabálycsoportokat törli [13][15]. A IV. stratégia [10] egy hierarchikus klaszterezési eljárással állapítja meg a lényegi szabályokat.

3. Szakértői tudásbázis beágyazása a FRIQ-learning módszerbe

Ha áll rendelkezésre valamilyen előzetes (a priori) tudásbázis a rendszer működésére vonatkozóan akkor célszerű ezen tudásbázis beépítése, majd az adott algoritmus futtatása ezen tudás alkalmazásával. Ebben az esetben, az előzetes tudásbázis helyességétől függően az adott megerősítéses tanuló algoritmus konvergencia sebessége nagymértékben javítható illetve az így kialakul modell jól használható olyan rendszerekben ahol az egzakt működés folyamata részben már ismert.

A cikk bemutat egy olyan módszert, amely által egy szakértő által megadott a priori tudásbázis beépíthető a FRIQ-learning megerősítéses tanulási módszerbe. Ez a szakértői heurisztika fuzzy szabályrendszer formájában építhető be, a FRIQ-learning rendszer tudásbázisát fuzzy szabályrendszer írja le. A rendszer kezdeti szabálybázisa fog kiegészülni a szakértő által definiált szabályokkal majd a tanulási folyamat közben az inkrementális szabálybázis építési módszer fogja hangolni azt.

A továbbiakban a szakértői heurisztika megadási módja, a rendszer kezdeti szabályrendszerével történő összefésülése, kezdeti Q-érték számítási módja a megadott szakértői heurisztikára és a rendszer működésének blokkvázlata kerül bemutatásra.

3.1. Szakértői heurisztika formája

Az FRIQ-learning tudásbázisa fuzzy szabályrendszer formájában áll elő. Ezek a szabályok az (1) formula alapján meghatározott módon épülnek fel, állapot-akció-Q-érték formájában, ahol az állapot-akció rész a szabály antecedense, a Q-érték pedig a konzekvense. A szakértői heurisztika fuzzy szabályok formájában adható meg, az *i*-edik ($i \in [1, r]$) szakértői szabály formája a következő:

$$\textbf{If } s_1 \textbf{ is } \hat{S}_1^i \textbf{ And } s_2 \textbf{ is } \hat{S}_2^i \textbf{ And … And } s_n \textbf{ is } \hat{S}_n^i \textbf{ Then } a = \hat{A}^i \tag{5}$$

Ez a formula részben hasonló az (1) formulához, azzal a kivétellel, hogy ezen szakértői szabályok esetében a szabály antecedensek az állapotok, a konzekvens pedig az ehhez az állapothoz tartozó megfelelő akció. Mivel Q-érték a szakértő által nehezen meghatározható (az egy, a rendszer számított érték a környezetből érkező megerősítések alapján) így ennek megadása nem lehetséges, további módszerre van majd szükség, amely azt meghatározza ezekre a szakértői szabályokra. Ennek következtében az (5) formulában az *i*-edik szakértői szabály konzekvens az $\hat{A}^i$ akció, az *n* dimenziós $\hat{\boldsymbol{S}}^i = \left[\hat{S}_1^i, \hat{S}_2^i, ..., \hat{S}_n^i\right]$ megfigyelt állapotban.

Ez a szakértői szabályrendszer egyben akcióválasztási politikát (π) is meghatároz a konzekvensenként megadott akciók következtében. Ha a megfigyelt állapotban áll rendelkezésre szakértő által megadott akció (azaz létezik rá szakértői szabály), akkor a rendszer ezt a megadott akciót fogja választani, a FRIQ-learning mohó (vagy ε-mohó) politikája által meghatározott akciója helyett. Ennek következtében ez a szabályrendszer egyben egy heurisztikus politika módosítóként [3] is tekinthető és az alábbi módon módosítja a FRIQ-learning mohó (vagy ε-mohó) politikáját:

$$\pi(\boldsymbol{s}) = \begin{cases} a = \hat{A}^i, & \text{if } \boldsymbol{s} = \hat{\boldsymbol{S}}^i, \textit{valamennyi } i-re \\ \arg\max\limits_{a \in U} Q^{\pi}(\boldsymbol{s}, a) & \textit{egyébként.} \end{cases} \tag{6}$$

ahol $\hat{\boldsymbol{S}}^i$ az n dimenziós állapota, $\hat{A}^i$ az ezen állapothoz tartozó akciója az i-edik szakértői szabálynak, s pedig az aktuális állapot megfigyelés. Ha az aktuális megfigyelés (**s**) illeszkedik valamelyik szakértői szabály antecedensére ($\hat{\boldsymbol{S}}^i$), akkor a rendszer által végrehajtott akció a szakértői által konzekvensként megadott akció ($\hat{A}^i$) lesz. Ellenkező esetben a rendszer által követett politika a mohó (vagy ε-mohó) lesz.

3.2. Kezdeti Q-érték számítása a szakértői szabályrendszerre

A szakértői szabálybázis formája az (5) formula által meghatározott felépítésű, ahol az antecedens az állapot, a konzekvens pedig az akció. Az FRIQ-learning módszer szabályrendszere állapot-akció-Q-érték formájú (1), ahol az antecedens az állapot-akció, a konzekvens pedig a Q-érték. Ennek következtében a szakértői szabályrendszerre szükséges valamilyen előzetes Q-érték meghatározása, hogy az adaptálható legyen a rendszerbe. Tehát a szakértői szabályok akció konzekvenseiből antecedenst kell előállítani, majd konzekvenseiként pedig előzetes Q-értéket szükséges meghatározni.

Feltételezve, hogy a szakértő által megadott szabályok megkérdőjelezhetetlenül helyesek, így az azokra meghatározott Q-értéknek relatívan magasnak kell lennie. A számított kezdeti Q-érték egy közelítés, amely a következő (7) összefüggés által határozható meg:

$$\tilde{Q}_{\text{init}} = \eta \cdot \frac{g_{\max}}{1-\gamma}, \text{ ha } \gamma < 1 \text{ esetében} \tag{7}$$

ahol $\tilde{Q}_{\text{init}}$ a számított kezdeti Q-érték, $g_{\max}$ a lehetséges maximális értékű megerősítés (konstans érték), $\eta \in [0,1]$ a $\tilde{Q}_{\text{init}}$értékre vonatkozó skála tényező, ami azt határozza meg, hogy az előzetesen számított $\tilde{Q}_{\text{init}}$érték mekkora részét (%-át) vegye figyelembe a rendszer.

Előfordulhat olyan eset mikor a megadott szabályok között található olyan is, amely nem feltétlenül helyes, azaz a szakértői heurisztika nem teljesen megfelelő, ebben az esetben az negatív hatással lehet a rendszer működésére.

3.3. Szakértői szabályrendszer beágyazása a FRIQ-learning kezdeti szabályrendszerébe

Az FRIQ-learning módszer 2^{n+1} darabszámú szabállyal rendelkező kezdeti szabálybázist hoz létre a tanulási folyamat kezdetével. Ezen szabályok konzekvense, azaz Q-értéke rendre $q_i = 0$ vesz fel, az *(n+1)*-dimenziós hiperkocka sarkaiban elhelyezkedve [12]. Ezen sarokponti szabályok formája a következő:

$$\textbf{If } s_1 \textbf{ is } S_1^{\square i} \textbf{ And } s_2 \textbf{ is } S_2^{\square i} \textbf{ And…And } s_n \textbf{ is } S_n^{\square i} \textbf{ And } a \textbf{ is } A^{\square i} \textbf{ Then } \tilde{Q}(\boldsymbol{s},a)=0 \tag{8}$$

ahol $S_l^{\square i} \in [\min(S_l), \max(S_l)], \forall i,l$ és $A^{\square i} \in [\min(A), \max(A)], \forall i$ a sarokponti állapot és akció értékek.

A szakértő által megadott heurisztikát leíró szabályok száma legyen $\hat{r}$. Mivel a rendszer kezdetben 2^{n+1} darabszámú szabállyal rendelkezik, így a teljes kezdeti szabályrendszer, tehát a FRIQ-learing kezdeti szabályai kiegészülve a szakértő által megadott szabályokkal, $2^{n+1} + \hat{r}$ darab szabályt fog tartalmazni. Ha a két szabályrendszer tartalmaz azonos szabályokat, akkor ez a szabálybázis méret csökken az egyező szabályok számával. A szakértői szabályok formája a kezdeti Q-érték meghatározási módszer által számított Q-értékekkel ($\tilde{Q}_{\text{init}}$) kiegészülve a következő:

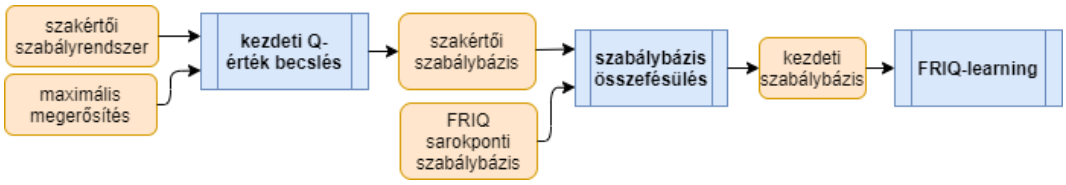
$$\textbf{If } s_1 \textbf{ is } \hat{s}_1^i \textbf{ And } s_2 \textbf{ is } \hat{s}_2^i \textbf{ And…And } s_n \textbf{ is } S_n^i \textbf{ And } a \textbf{ is } \hat{A}^i \textbf{ Then } \tilde{Q}(\boldsymbol{s},a)=\tilde{Q}_{\text{init}},\ i \in [1,\hat{r}] \tag{9}$$

A (9) formulájú szakértői szabályrendszer FRIQ-learning módszerbe történő beágyazásához össze kell fésülni ezen szabályokat a (8) formulával leírt sarokponti szabályokkal. A két szabályrendszer összefésülésekor előfordulhat olyan eset mikor egy szakértő által megadott szabály éppen valamelyik sarokponti szabályra esik. Ilyenkor ellentmondás lép fel, mert az illeszkedő szabályok antecedense ugyanaz, de a konzekvensük különböző. Ekkor a sarokponti szabály következménye $\tilde{Q}(\boldsymbol{s},a)=0$ a szakrétő szabályé pedig $\tilde{Q}(\boldsymbol{s},a)=\tilde{Q}_{\text{init}}$, ahol $\tilde{Q}_{\text{init}}$nagy valószínűséggel 0-től különböző érték. Ezt az ellentmondásos állapotot fel kell oldani. Ez olyan módon valósul meg, hogy a kifejlesztett módszer a két szabályrendszer összefésülésekor ellenőrzi, hogy valamely szakértői szabály sarokponti szabályra esik-e. Ha igen, akkor lecseréli a sarokponti szabályt a szakértői szabályra, azaz törli a sarokponti szabályt, majd a sarokponti szabály konzekvense a szakértőire számolt Q-értéket fogja felvenni. Az összefésült szabályrendszer szabályainak száma ($2^{n+1} + \hat{r}$) csökkeni fog az egyező szabályok számával.

Mivel a szakértő által megadott szabályok következmény része valószínűleg 0-tól különböző érték lesz, így az jelentősebb befolyással van a rendszer működésére, ezek a szabályok nagyobb súllyal kerülnek figyelembevételre.

3.4. A rendszer felépítése

A szakértői heurisztika adoptálásának módját és az így létrejött szakértői tudás alapú FRIQ-learning rendszer felépítését a következő szemlélteti:



1. ábra. *Szakértői heurisztika adoptálásának módja az FRIQ-learning rendszerbe*

Az ábrán látható, hogy a rendszer bemenete a megadott szakértői tudásbázis és a környezet által adható lehető legnagyobb megerősítés. A megadott a priori szabályokra kezdeti Q-érték számítása történik a (7) összefüggéssel meghatározott módon. Ezt követően a már Q-értékekkel rendelkező szakértői szabálybázis és a FRIQ-learning rendszer sarokponti szabálybázisa kerül összefésülésre a 3.3-as alfejezetben leírtak szerint. Az így létrejött kezdeti tudásbázis kerül beillesztésre az FRIQ-learning rendszerbe. Az adoptálást követően a FRIQ-learning tanulási fázisa következik, az inkrementális szabálybázis építési (és hangolási) [12] módszer alkalmazásával.

4. Mountain car mintapélda szakértői tudásbázis alkalmazásával

Szakértői heurisztika FRIQ-learning rendszerbe történő beillesztését, a kifejlesztett szakértői tudásbázis adoptálását egy elterjedt megerősítéses tanulási mintapéldán keresztül mutatja be ezen fejezet.

A választott mintaalkalmazás a népszerű megerősítéses tanulási problémák közül a „mountain car” nevezetű. Ebben az esetben az ágens egy autó, környezete pedig egy meredek völgy. Az autó a meredek völgy közpén helyezkedik el a tanulási folyamat indulásakor. A ágens célja, hogy kijusson a meredek völgy közepéből a völgy tetején található dombra. A feladat akkor tekinthető megoldottnak, ha az autó valamennyi meghatározott lépés alatt (jelen esetben 1000) kijut a völgyből. Ebben az esetben a környezettől egy nagy megerősítést kap, ellenkező esetben büntetést. A „mountain car” probléma állapot tere 2 változós. Ezek az autó aktuális pozícióját (s_1) és sebességét (s_2) leíró állapotváltozók. Az akció tér 1 dimenziós, amely az autó elmozdulását (a) írja le. Ez a következő értékeket veheti fel: jobbra, balra vagy nincs elmozdulás.

A rendszer szakértői tudásbázisának megadása jelen esetben fuzzy szabályok formájában lehetséges. Jövőbeli kutatási tervek között szerepel, hogy a szakértői heurisztika megadása egy leírónyelv [6] alkalmazásával valósuljon meg, amely az emberi gondolkodáshoz közelebb álló megadási formát tesz lehetővé.

A szakértői által fuzzy szabályok formájában megadott heurisztika rendszerre gyakorolt hatásának megállapításához 4 futtatási esetet hoztunk létre. Első esetben egy helyesen megadott heurisztikával, második esetben az előzőekben megadott heurisztika csak egy részével, harmadik esetben részben helytelen szakértői szabályrendszerrel, negyedik esetben pedig egy „véletlenszerűen” generált szakértői szabálybázissal került futtatásra a mintapélda. Az összehasonlítás alapja az üres tudásbázissal, azaz a szakértői heurisztika nélkül működő FRIQ-learning rendszer által adott konvergencia sebesség és fuzzy szabálybázis méret (szabályszám). Minden egyes esetben 10 külön, egymástól független futtatás valósult meg, különböző kezdeti állapottér pozíciókkal. Az egyes futtatási eredmények adataiból, az egyes konvergencia sebességek és szabálybázis méretek átlagai kerültek meghatározásra. A szabály-

bázis méretek mindegyik esetben a szabálybázis redukálási stratégiák alkalmazása nélküli szabályszámokat jelölik.

A rendszer futtatási paraméterei a következők:

- szakértő által megadott megerősítés g_{max}=100
- tanulási ráta α=0.5
- leszámítolási tényező γ=0.99

Amikor a rendszer szakértői tudásbázis nélkül kerül futtatásra akkor az átlagos konvergencia sebesség 28.3 epizód, szabálybázis méret pedig 91.7 szabály. A következő táblázat foglalja össze az ebben az esetben kapott futási eredményeket.

1. táblázat. *Szakértői heurisztika nélküli futtatási eset eredményei*

Futtatási eset	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Átlag
Konvergencia sebesség	23,36,34,35,20,34,25,26,29,21	28.3
Szabálybázis méret	80,85,82,96,105,90,89,98,99,93	91.7

A következő futtatási esetben a helyesen megadott szakértői tudásbázis beillesztésre kerül a rendszerbe és a tanulási fázis ezzel a szabálybázissal indul el. A helyes szakértői szabálybázist az előző futtatási esetből, az inkrementálisan felépített szabálybázisból nyertük ki a szabálybázis redukálási módszerek által. Az így kapott 17 szabályt a következő táblázat tartalmazza.

2. táblázat. *A helyesen megadott szakértői szabályok felépítése*

R#	**s_1**	**s_2**	**a**
1	-0.5	0	-1
2	-0.475	-0.014	1
3	-0.475	0.014	1
…	…	…	…
15	-0.65	0.042	0
16	-1.09	0.042	-1
17	0.14	-0.014	0

A maximális szakértő által megadott megerősítés (g_{max}=100) következtében a 2. táblázatban lévő szabályokra a (7) összefügés által számított kezdeti Q-értékek $\tilde{Q}_{\text{init}} = 10000$.

A következő lépésben a (7) összefüggésben lévő η érték jelentőségének vizsgálata valósult meg. Tehát, hogy a lehetséges $\tilde{Q}_{\text{init}}$érték valamekkora részét figyelembe véve hogyan változik a konvergencia sebesség. A kapott eredményeket a következő táblázat tartalmazza:

A 3. táblázatban lévő futási eredményekből az látható, hogy a $\tilde{Q}_{\text{init}}$érték egyre kisebb részét figyelembe véve egyre jobban romlik a rendszer konvergencia sebessége. A kapott adatok alapján jelen esetben a $\tilde{Q}_{\text{init}}$érték 100%-a kerül figyelembe vételre, azaz η=1.

3. táblázat. *Az η érték konvergencia sebességre gyakorolt hatása*

η	**Konvergencia sebesség (epizódok száma)**	$\tilde{Q}_{\text{init}}$
100	23	10000
75	23	7500
60	30	6000
37	29	3700
7.5	25	750
0.015	27	1.5

A helyesen megadott szakértői heurisztikával történő futtatás eredményeit a következő 4. táblázat tartalmazza. Ebben az esetben a rendszer átlagosan 10 epizód alatt és 124.3 szabálybázis mérettel találta meg a megoldást.

4. táblázat. *Helyes szakértői heurisztikával történő futtatás eredményei*

Futtatási eset	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Átlag
Konvergencia sebesség	10,20,17,7,11,10,6,5,6,8	10
Szabálybázis méret	108,125,139,109,135,129,107,124,133,134	124.3

A következő esetben a helyesen megadott szakértői szabályrendszer egy részével indult el a tanulási fázis. Ebben az esetben az előzőekben megadott 2. táblázatban lévő 17 szakértő szabályból kiemeltünk néhány darabot, véletlenszerűen, pontosan a 10 darabot a 17-ből. Az így kapott szabályrendszer helyes, de kisebb méretű, mint az előzőekben megadott szabályrendszer (2. táblázat). Ebben az esetben átlagosan 14.4 epizód alatt és 114.3 szabályszámmal konvergált a rendszer. A pontos futási eredményeket a következő 5. táblázat tartalmazza.

5. táblázat. *A helyes szakértői heurisztika egy részével történő futtatás eredményei*

Futtatási eset	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Átlag
Konvergencia sebesség	20,13,10,7,7,15,29,15,22,6	14.4
Szabálybázis méret	107,85,102,85,98,96,111,107,110,98	114.3

A következő esetben azt vizsgáltuk, hogy milyen hatással van a rendszerre az, ha a helyesnek feltételezett szakértői szabályrendszer tartalmaz néhány „rossz” szabályt is. A rossz szabályok azt jelentik, hogy az adott szakértői szabály antecedenshez nem megfelelő konzekvens lett meghatározva. Ez azt jelentheti, hogy az ágens által végrehajtott cselekvéssorozat elromlik abban az értelemben, hogy az ágens ennek következtében nem fog eljutni a célállapotba. A 17 helyesen definiált szakértői szabályrendszerből 6 szabály konzekvensét elrontottuk úgy, hogy módosítottuk az adott akciót. Ezt a szabályrendszert a következő táblázat tartalmazza, amelyben csak az elrontott szabályokat tüntettük fel, ezek sorszáma a következő: 1, 2, 3, 15, 16 és 17.
Az így kapott futtatási eredményeket a 7. táblázat tartalmazza, ebben az esetben átlagosan 11.7 epizóddal és 120.1 szabályszámmal konvergált a rendszer. A futási eredményeket a következő 5. táblázat tartalmazza:
Az utolsó futtatási eset mikor „véletlenszerűen” generált szakértői heurisztikával indul a rendszer. Ebben az esetben szintén 17 szabályból áll a szakértői szabályrendszer, de ezek véletlenszerű állapotokkal (antecedenssel) és akcióval (konzekvenssel) rendelkeznek. Ezen állapot- és akció értékek vélet-

lenszerűen lettek létrehozva, adott tartományon belül. Ezen szabályok közül néhányat a 8. táblázat tartalmaz.

6. táblázat. *A részben helyes szakértői szabályrendszer helytelen szabályai*

R#	**s_1**	**s_2**	**a**
1	-0.5	0	0
2	-0.475	-0.014	1
3	0.475	-0.014	-1
...	…	…	…
15	-0.68	0.042	0
16	-1.09	0.042	0
17	0.14	-0.014	1

7. táblázat. *Helyes szakértői heurisztika egy részével történő futtatás eredményei*

Futtatási eset	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Átlag
Konvergencia sebesség	8,16,8,13,7,16,10,15,16,7	11.7
Szabálybázis méret	115,134,126,133,135,126,123,135,147,127	120.1

8. táblázat. *A „véletlenszerűen” generált helytelen szakértői heurisztika*

R#	**s_1**	**s_2**	**a**
1	-0.475	0	1
2	-0.5	0	-1
3	-0.475	-0.014	-1
...	…	…	…
15	0.885	0.042	1
16	-0.065	0.042	0
17	-1.09	0.042	-1

Ezen szabályrendszerrel történő futtatás eredményeit a 9. táblázat tartalmazza.

9. táblázat. *A „véletlenszerűen” generált szakértői heurisztikával történő futtatás eredményei*

Futtatási eset	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Átlag
Konvergencia sebesség	29,56,19,16,24,18,37,29,20,17	26.6
Szabálybázis méret	122,127,118,124,131,120,130,124,127,121	124.4

Összegzésként a 10. táblázat tartalmazza az egyes futtatási esetek eredményeit.

10. táblázat. *Az egyes futtatási estek eredményei összegezve*

Szakértői heurisztika típusa	Átlagos konvergencia sebesség	Átlagos szabálybázis méret
Üres (heurisztika nélkül)	28.3	91.7
Helyesen megadott	10	124.3
Helyesen megadottnak egy része	14.4	114.3
Részben helytelenül megadott	11.7	120.1
Véletlenszerűen generált	26.6	124.4

5. Összefoglalás

A cikkben bemutatott eljárás segítségével a FRIQ-learning megerősítéses tanulási módszerbe szakértő által megadott heurisztika illeszthető. Bemutatásra került az előzetes szakértői tudásbázis leírásának módja és egy előzetes Q-érték számítási módszer is, amely a szakértő által megadott megerősítési érték alapján Q-értékeket határoz meg szakértői szabályokra, hogy azok illeszthetők legyenek a rendszerbe. Bemutatásra került továbbá egy szabálybázis összefésülési módszer, amely a FRIQ-learning kezdeti sarokponti szabályait a szakértői által megadott szabályrendszerrel összeolvasztja, majd ezzel az összefésült szabályrendszerrel indítja el a tanulási folyamatot.

A kifejlesztett előzetes szakértői tudásbázis adoptálásának szemléltetésére és a szakértői heurisztika rendszerre gyakorolt hatásának bemutatására az elterjedt „mountain car” mintapéldát választottuk. 4 különböző futtatási esetet hoztunk létre. Mindegyik eset 10 egymástól független futtatással valósult meg majd az így kapott futási eredmények átlagait határoztuk meg. Első esetben egy helyesen megadott heurisztikával, második esetben az előzőekben megadott heurisztika csak egy részével, harmadik esetben részben helytelen szakértői szabályrendszerrel, negyedik esetben pedig egy „véletlenszerűen” generált szakértői szabálybázissal került futtatásra a mintapélda. Az összehasonlítás alapja a szakértői heurisztika nélkül működő FRIQ-learning rendszer konvergencia sebessége és fuzzy szabálybázisban lévő szabályok száma volt. Helyesen megadott szakértő heurisztika javította a rendszer konvergencia sebességét, átlagosan 10 epizódra, 28.3 epizódról. A helyes szakértői tudásbázisnak csak egy részével és a részben helytelenül megadott szakértői tudással történő futtatások rontottak a konvergencia sebességen, átlagosan 14.4 és 11.7 epizód, de ezek az értékek még mindig jobbak, mint ha heurisztika nélkül (28.3 epizód) futna a rendszer. A véletlenszerűen generált szakértői tudásbázis rontotta a rendszer hatékonyságát, átlagosan 26.6 epizód alatt oldotta meg a problémát. Ez az érték rosszabb, mint a heurisztika nélküli futtatási eset eredményei, tehát a helytelen előzetes tudásbázis negatív hatással van a rendszerre.

Az eredmények alapján elmondható, hogy egy helyesen megadott, előzetes szakrétői tudásbázis nagymértékben javítja a FRIQ-learning rendszer konvergencia sebességét, pozitív módon befolyásolja a tanulási folyamatot.

Jövőbeli cél egy gradiens módszeren alapuló szabálybázis hangolási eljárás kidolgozása, amely a megadott szakértői heurisztika hangolására, tehát a fuzzy szabályok pozíciójának elmozgatására képes a többdimenziós állapot-akció-Q-érték térben.

Irodalom

[1] Appl, M.: Model-based Reinforcement Learning in Continuous Environments. Ph.D. thesis, Technical University of München, München, Germany, dissertation.de, Verlag im Internet (2000)

[2] Berenji, H.R.: Fuzzy Q-Learning for Generalization of Reinforcement Learning. Proc. of the 5th IEEE International Conference on Fuzzy Systems (1996) pp. 2208-2214.

[3] Bianchi, Reinaldo AC, Carlos HC Ribeiro, and Anna HR Costa. "Accelerating autonomous learning by using heuristic selection of actions." Journal of Heuristics 14.2 (2008): 135-168. **https://doi.org/10.1007/s10732-007-9031-5**

[4] Broekens, Joost, Koen Hindriks, and Pascal Wiggers. "Reinforcement learning as heuristic for action-rule preferences." International Workshop on Programming Multi-Agent Systems. Springer Berlin Heidelberg, 2010.

[5] Kovács, Sz.: "Extending the Fuzzy Rule Interpolation "FIVE" by Fuzzy Observation", Computational Intelligence, Theory and Applications: 9th International Conference on Dortmund Fuzzy Days. 802 p., Dortmund, Germany, 2006.09.18-2006.09.20. Berlin Heidelberg: Springer-Verlag, 2006. pp. 485-497.

[6] Piller, Imre, and Szilveszter Kovács. "Fuzzy Behavior Description Language: A Declarative Language for Interpolative Behavior Modeling." Acta Polytechnica Hungarica 16.9 (2019). **https://doi.org/10.1109/INES46365.2019.9109451**

[7] Pourhassan, Mojgan, and Nasser Mozayani. "Incorporating expert knowledge in Q-learning by means of fuzzy rules." Computational Intelligence for Measurement Systems and Applications, 2009. CIMSA'09. IEEE International Conference on. IEEE, 2009. **https://doi.org/10.1109/CIMSA.2009.5069952**

[8] Sutton, R. S., Barto, A. G.: Reinforcement Learning: An Introduction, MIT Press, Cambridge (1998) **https://doi.org/10.1109/TNN.1998.712192**

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

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

[11] Vincze, D., Kovács, Sz.: "Fuzzy rule interpolation-based Q-learning." Applied Computational Intelligence and Informatics, 2009. SACI'09. 5th International Symposium on. IEEE, 2009. **https://doi.org/10.1109/SACI.2009.5136311**

[12] Vincze, D., Kovács, Sz.: Incremental Rule Base Creation with Fuzzy Rule Interpolation-Based Q-Learning, I. J. Rudas et al. (Eds.), Computational Intelligence in Engineering, Studies in Computational Intelligence, Volume 313/2010, Springer-

Verlag, Berlin Heilderberg, 2010, pp. 191-203. **https://doi.org/10.1007/978-3-642-15220-7_16**

[13] Vincze, D., Kovács, Sz.: Reduced Rule Base in Fuzzy Rule Interpolation-based Q-learning, Proceedings of the 10th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, CINTI 2009, November 12-14, 2009, Budapest Tech, Budapest, pp. 533-544. **https://doi.org/10.1109/SACI.2009.5136311**

[14] Vincze, D.: Fuzzy Rule Interpolation and Reinforcement Learning, 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI 2017), Herl'any, Slovakia, pp. 173–178. **https://doi.org/10.1109/SAMI.2017.7880298**

[15] Vincze, D., Kovács, Sz.: Rule-Base Reduction in Fuzzy Rule Interpolation-Based Q-Learning, Recent Innovations in Mechatronics (RIiM) Vol. 2. (2015) No. 1-2. **https://doi.org/10.17667/riim.2015.1-2/10.**

[16] Watkins, C. J. C. H.: Learning from Delayed Rewards. Ph.D. thesis, Cambridge University, Cambridge, England (1989)