Production Systems and Information Engineering
Volume 10 (4), pp. 37–47.
http://doi.org/10.32968/psaie.2022.4.4

TUDÁSBÁZIS HANGOLÁSA A FRIQ-LEARNING MEGERŐSÍTÉSES TANULÁSI 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 klasszikus megerősítéses tanulási rendszerekben a probléma megoldását leíró tudásbázis ismeretlen a tanulási folyamat kezdetén. Ezen módszerek többsége próbálkozás alapú keresést valósít meg, a környezet visszajelzései alapján térképezi fel a lehetséges megoldást. Azonban, ha rendelkezésre áll részinformáció a probléma megoldására vonatkozóan és az adaptálható a rendszerbe, akkor a tanulási folyamat hatékonysága javítható. A szakértői tudásbázissal bővített Fuzzy-szabályinterpoláció alapú Q-tanulás (expert knowledge-included Fuzzy Rule Interpolation-based Q-learning) rendszerben előzetes szakértői információ (szakértői tudásbázis) állapot-akció típusú Fuzzy-szabályok formájában injektálható a rendszer tanulásfolyamatába, amely által a módszer konvergenciasebessége javítható. Azonban, abban az esetben, ha az előzetes szakértői tudásbázis helytelen információkat tartalmaz a megoldásra vonatkozóan, akkor ez negatív hatással lehet a tanulási folyamat hatékonyságára. A cikk célja, egy olyan javasolt hangolási (optimalizálási) eljárás bemutatása, amely a tanulási folyamat során alkalmas lehet a helytelen információkat leíró szakértői Fuzzy-szabályrendszer hangolására, azaz a Fuzzy-szabályok állapot-akció pontjának optimalizálására.

Kulcsszavak: megerősítéses tanulás, heurisztikusan gyorsított megerősítéses tanulás, szakértői tudásbázis, tudásbázis hangolás, Q-learning, Fuzzy Q-learning

1. Bevezetés

A megerősítéses tanulás (Reinforcement Learning – RL) [11] olyan gépi tanulási módszer, amely működése a környezet által adott visszajelzéseken (megerősítéseken) alapszik. Ezen próbálkozás típusú módszerek az ágens teljesítményét (hatékonysá-

gát) annak szerzett tapasztalatai által igyekeznek javítani, törekedve arra, hogy a gyűjtött jutalmakat hosszú távon maximalizálja.

A klasszikus megerősítéses tanulási módszerek (pl. Q-learning [2], Fuzzy Q-learning [1] és SARSA [6]) a tanulási folyamat előtt nem rendelkeznek információról az adott probléma megoldására vonatkozóan, majd a kezdeti üres tudásbázisukat a tanulási folyamat során töltik fel (és finomítják) iterációról-iterációra. Ezen módszerek általános célja a probléma megoldását leíró Q-függvény (állapot-akció értékfüggvény) keresése.

A tudásbázis reprezentáció és így a Q-függvény leírásának módja RL módszerenként eltérő lehet, Q-learning- és SARSA-módszerek esetében ez egy Q-tábla (többdimenziós mátrix), Fuzzy-modell alapú RL-módszerek esetében pedig „ha-akkor" típusú Fuzzy-szabályokból álló szabálybázis.

A „Heurisztikusan Gyorsított Megerősítéses Tanulás" ("Heuristically Accelerated Reinforcement Learning" – HARL) [10] olyan RL módszerek, amelyek lehetőséget adnak külső információ injektálására a rendszer tudásbázisába. Ezekben az esetekben egy heurisztikus függvény írja le a külső információt, amely meghatározza az ágens számára az adott állapotokban javasolt akciókat.

A szakértői tudásbázissal bővített Fuzzy-szabályinterpoláció alapú Q-tanulás (expert knowledge-included Fuzzy Rule Interpolation-based Q-learning) rendszer [24] lehetőséget ad külső szakértő által meghatározott információ rendszerbe történő beágyazásra. A rendszer tudásbázisa (Q-függvénye) egy ritka Fuzzy-szabálybázis által leírt, a szakértői előzetes tudás pedig állapot-akció típusú Fuzzy-szabályok formájában definiálható [15] [18]. A megoldásra vonatkozó információ valamely része ezen szakértő által definiált állapot-akció formájú szabályok által írható le. Ha az a priori tudásbázis helyes információkat tartalmaz a megoldásra vonatkozóan, akkor ebben az esetben a rendszer konvergenciasebessége (betanuláshoz szükséges epizódok, lépések száma) javulhat [15][20]. Abban az esetben, ha a szakértői heurisztika helytelen információt tartalmaz, azaz az adott állapothoz nem megfelelő akcióérték társul (helytelen döntés), akkor az a rendszer tanulási folyamatára negatív hatással lehet, a konvergenciasebessége nagymértékben lassulhat [15] [20]. Ennek következtében szükséges egy módszer, amely alkalmas a tudásbázist leíró szabálybázis (a szakértői szabályokat is beleértve) hangolására (optimalizálására).

Számos optimalizálási módszer [9] található a szakirodalomban (például gradiens módszer [7] [12], részecske-raj alapú optimalizálás [8] stb.), amely alkalmas lehet az RL-rendszer paramétereinek optimalizálására. Például a gradiens módszer alapú optimalizálást a neurális hálózat súlyainak hangolására a Deep Q-learning Network (DQN) alkalmaz [23].

Jelen cikk célja egy javasolt hangolási (optimalizálási) eljárás bemutatása, amely alkalmas lehet a FRIQ-learning szabálybázisát alkotó szabályok (beleértve a szakértői szabályokat) antecedens és konzekvens értékeinek tanulási folyamat közbeni hangolására.

2. Szakértői heurisztikával bővített FRIQ-learning

A szakértői tudásbázissal bővített FRIQ-learning [15] [18] az FRIQ-learning (Fuzzy Rule Interpolation (FRI) based Q-learning) módszer [24] kibővítése, amely lehetőséget ad külső szakértő által leírt információ (a priori tudásbázis) FRIQ-learning rendszerbe történő injektálására. A FRIQ-learning módszer a „FIVE" FRI [13] Fuzzy-szabályinterpolációs modellt alkalmazza a Q-függvény leírására, amely következtében a rendszer működtető tudásbázisát egy ritka Fuzzy-szabálybázis írja le. Egyetlen $r_i(i \in [1,m], m\text{: } szabályszám)$ szabály formája az R jelölésű szabálybázisban a következő [25]:

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

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 által becsült Q-függvény, q^i pedig az i-edik szabály konzekvense (Q-értéke).

A szakértői tudásbázis formája az (1) összefüggés által meghatározott szabályok formájához hasonló, de azzal az eltéréssel, hogy ebben az esetben az antecedens az állapot, a konzekvens pedig az ebben az állapotban preferált akció [15]:

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

ahol $\hat{S}_n^i = \left[\hat{S}_1^i, \hat{S}_2^i, \dots \hat{S}_n^i\right]$ az n-dimenziós állapotmegfigyelés, $\hat{A}^i$ az ehhez az állapotmegfigyeléshez tartozó akció, i ($i \in [1,\hat{m}]$) pedig a szabály sorszáma az $\hat{m}$ méretű szakértői szabálybázisban.

A FIVE Fuzzy-szabályinterpolációs módszer alkalmazása következtében a szakértői szabályrendszer tetszőleges darabszámú szabályt, bármilyen állapot és akció pontban tartalmazhat, definiálva általa az ágens viselkedését (azaz az adott állapotban a preferált akció végrehajtását).

A rendszer tanulási folyamata a szakértői szabályrendszer [15] és az n-dimenziós hiperkocka sarkaiban elhelyezkedő sarokponti szabályok [25] inicializálásával kezdődik. A 0 konzekvens értékkel rendelkező, 2^{n+1} (n: állapotdimenziók száma) darabszámú $r_i^{\square}$ sarokponti szabály a FIVE Fuzzy-szabályinterpolációs módszer alkalmazása által szükséges. Ezen szabályok formátuma a következő:

$$r_i^{\square}\text{: } \boldsymbol{If}\ s_1\ is\ S_1^{\square i}\ \boldsymbol{And}\ s_2\ is\ S_2^{\square i}\ \boldsymbol{And}\ \dots\ \boldsymbol{And}\ s_n\ is\ S_n^{\square i}\ \ \boldsymbol{And}\ a\ is\ A^{\square i}\ \boldsymbol{Then}\ \tilde{Q}(\boldsymbol{s},a) = 0 \qquad (3)$$

A szakértői szabályrendszer injektálása során az állapot-akció formátumú szakértői szabályok (2) módosulnak az (1) összefüggésnek megfelelően. A szakértői szabá-

lyok akciókonzekvense antecedensre módosul, majd az új konzekvensük egy becsült $\tilde{Q}_{init}$ érték lesz, amely egy javasolt Q-érték becslési módszer [15] által kerül meghatározásra (a Q-érték becslési módszerről további információk a [15] sorszámú hivatkozásban):

$$\hat{r}_i\text{: } \textbf{If } s_1 \textbf{ is } \hat{S}_1^i \textbf{ And } s_2 \textbf{ is } \hat{S}_2^i \textbf{ And } \dots \textbf{And } s_n \textbf{ is } \hat{S}_n^i \textbf{ And } \text{a} = \hat{A}^i \textbf{ Then } \tilde{Q}(\boldsymbol{s}, a) = \tilde{Q}_{init} \qquad (4)$$

A szakértői és a sarokponti szabályokat tartalmazó két szabálybázis összefésülésre kerül egyetlen szabálybázissá olyan módon, hogy a szakértői szabályokra esetlegesen illeszkedő (és így ellentmondó) sarokponti szabályok 0 értékű konzekvense lecserélésre kerül a szakértői szabály konzekvensére, majd a sarokponti szabály törlésre kerül (feloldva az ellentmondást).

A tanulási folyamat során az így létrejött kezdeti szabálybázis, amely a szakértői és a sarokponti szabályokat tartalmazza, inkrementálisan fog növekedni új szabályok hozzáadásával [25]. Új szabály akkor kerül beszúrásra a kezdeti szabályrendszerbe, ha Q-frissítés értéke nagyobb, mint egy küszöbérték ($\Delta\tilde{Q} > \varepsilon_Q$) és az aktuális megfigyeléshez legközelebb elhelyezkedő szabály is távolinak [22] tekinthető. Az új szabályok állapot-akció pozíciója az aktuális megfigyelés állapot-akció pontja lesz. Ha a megfigyelés közelében helyezkedik el már létező szabály és a Q-frissítés értéke is relatívan kicsi, akkor a már létező szabályok konzekvense (Q-értéke) kerül frissítésre az alábbi összefüggés által [24] [25]:

$$\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \tilde{Q}^k(\boldsymbol{s}, a) + \Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) \qquad (5)$$

$$\Delta\tilde{Q}^{k+1}(\boldsymbol{s}, a) = \alpha * \left(g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a') - \tilde{Q}^k(\boldsymbol{s}, a)\right) \qquad (6)$$

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 $\boldsymbol{s} \to \boldsymbol{s}'$ állapotátmenetre, $\tilde{Q}^k$ és $\tilde{Q}^{k+1}$ pedig a k-adik és a $(k+1)$-edik iteráció FIVE FRI módszer által becsült Q-értéke [24]:

$$\tilde{Q}(\boldsymbol{s}, a) = \begin{cases} \sum_{i=1}^{m} \left(\left(q^i \Big/ (\delta_v^i)^\lambda \right) \Big/ \left(\sum_{j=1}^{m} 1 \Big/ (\delta_v^j)^\lambda \right) \right) & \text{ha } (\boldsymbol{s}, a) = (\boldsymbol{s}^i, a^i) \text{ valamennyi } i - re, \\ q^i & \text{egyébként} \end{cases} \qquad (7)$$

ahol q^i az i-edik ($i \in [1, m]$) szabály konklúziója, (s, a) a megfigyelés, δ_v^i a skálázott távolság [13] az (s, a) megfigyelés és az i-edik szabály (s^i, a^i) antecedense között, λ a Shepard-paraméter, m pedig a szabályok száma (további információk a [13], [15], [24] és [25] sorszámú hivatkozásokban). A δ_v^i skálázott távolság [13] a következőképpen határozható meg:

$$\delta_v^i = \delta_v\left((\boldsymbol{s}, a), (\boldsymbol{s^i}, a^i)\right) = \left[\sum_{j=1}^{n}\left(\int_{s_j^i}^{s_j} c_j\, dx_j\right)^2 + \left(\int_{a^i}^{a} c_a dx_a\right)^2\right]^{1/2} \qquad (8)$$

ahol $(\boldsymbol{s}, a)$ az állapot-akció megfigyelés, $(\boldsymbol{s^i}, a^i)$ az i-edik szabály állapot-akció antecedense, s_j a j-edik ($j \in [1, n]$) dimenziója az n-dimenziós állapottér univerzumnak, s_j^i az i-edik szabály j-edik állapotdimenziója, a^i az i-edik szabály akcióuniverzuma, c_j az s_j állapot univerzum konstans skálafüggvénye, c_a pedig az U akcióuniverzum konstans skálafüggvénye.

A tanulási fázis végeztével szabálybázis-redukálási módszerek [3] [4] [16] [26] alkalmazhatóak az inkrementálisan létrejött szabálybázis elhagyható szabályainak keresésére. Ezen szabálybázis-redukálási módszerek által a Q-függvényt leíró szabálybázis mérete csökkenthető.

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

Az előzőekben bemutatott FRIQ-learning módszer a tudásbázist leíró Fuzzy-szabályok konzekvensének (azaz a Q-értékének) hangolja a (6) frissítési formula alapján, azonban az újonnan felvett szabályok antecedens része változatlan marad a teljes tanulási folyamat során. Abban az esetben, ha a rendszer vesz fel új szabályokat a szabálybázisba, akkor az újonnan létrehozott szabály állapot-akció antecedense (szabálypont) az állapot-akció térrácsháló [25] az aktuális állapot-akció értékéhez legközelebbi pontjába kerül.

A javasolt szakértői tudásbázissal bővített rendszer esetében az eredetileg alkalmazott állapot-akció térrácsháló [25] elhagyásra került, ezért az újonnan létrehozott szabályok antecedense pontosan az aktuális megfigyelés állapot-akció pontjába kerül. Ennek következtében a szabályok nem a rácsháló által meghatározott pontokban, hanem tetszőleges állapot-akció pontokban helyezkedhetnek el. A szakértő által definiált a priori szabályrendszer esetében, amennyiben a megadott szakértői produkciós szabály valamelyik állapothoz nem megfelelő akcióértéket rendel (azaz a szakértő szabályrendszer csak részben tekinthető helyesnek), úgy a szabálybázisba felvett szakértői szabály állapot-akció antecedense is rossz helyre kerül. Ebben az esetben, a csak részben helyes szabályrendszer negatív hatással lehet a tanulási folyamat hatékonyságára [15] [20], így szükség lehet egy hangolási eljárásra, amely képes a szabályok állapot-akció pontját, azaz antecedensét elhangolni (optimalizálni) a megfelelő irányba, szabálypontba. Fennállhat olyan eset is, amikor akár egy teljes antecedens dimenzió is szükségtelen lehet, ezen antecedenes redundancia felderítése történhet utólagosan is [3].

A fentiek alapján a javasolt szakértői tudásbázissal bővített FRIQ-learning (HFRIQ-learning) módszer az alábbi főbb lépésekből áll:

- A tanulási fázis, azaz a szabálybázis létrehozási folyamat során a rendszer kezdeti (sarokponti és szakértői) szabálybázisa kiegészül a rendszer által felvett új szabályokkal.
- Ha az éppen vizsgált állapot-akció (megfigyelés) pontban még nem létezik szabály és a legközelebbi szabály is távolinak számít, akkor a rendszer felvesz ebbe a pozícióba egy új szabályt, amely állapot-akció pontja megegyezik az éppen aktuális megfigyelés állapot-akció pontjával.
- Új szabály felvételekor a szabályok közötti közelségmérték és egy megengedett minimális szabálytávolság meghatározása, mely alapján két szabály egymáshoz közelinek tekinthető.
- A szabályok közötti távolság számítása antecedens dimenziónként. Két szabály akkor tekinthető közelinek, ha minden egyes antecedens univerzumban közeliek [22].
- Ha az éppen vizsgált állapot-akció pontban, vagy ahhoz közel már van létező szabály, akkor a szabálybázis szabályai hangolódnak, azaz a szabálypontok vándorolnak (gradiensalapú optimalizációs módszer esetén a Q-függvény gradiensének megfelelően).
- Ha a hangolási folyamat során két szabály közel kerül egymáshoz a szabályvándorlás következtében, akkor ezen szabályok egyetlen kardinális szabályként egyesülnek (csökkentve a szabálybázis méretét már a tanulási fázis során) [19].

3.1. A gradiens módszer alkalmazása a szabályrendszer hangolására

Abban az esetben, ha az éppen aktuális megfigyelés közelében már található létező Fuzzy-szabály, akkor nem kerül új szabály felvételre a megfigyelés pozíciójába, hanem a szabályrendszer antecedense és konzekvense kerül hangolásra a gradiens módszer alkalmazásával.

A gradiens módszer egy iteratív optimalizációs algoritmus, amely célja egy F függvény minimumpontjának meghatározása, amelyet úgy valósít meg, hogy egy adott x_0 függvénypontból kiindulva iterációról iterációra valamekkora α lépést megtéve halad a legmeredekebb lejtő irányába, amely irányt a függvény változói szerinti deriváltjai (iránymenti deriváltjai) határozzák meg. A módszer alkalmazhatóságának feltétele (a gradiens számítása következtében) az adott függvény differenciálhatósága. Többváltozós függvény esetében minden egyes változóra szükséges a parciális deriváltak meghatározása. A parciális deriváltakból képzett ∇ vektor a gradiens vektor, amely egy adott függvénypontban megmutatja, hogy merre növekszik a függvény a leginkább. Az F függvény minimumpontjának keresése során egy x_k pontból kiindulva az iteráció rákövetkező x_{k+1} értéke úgy áll elő, hogy az egy α tanulási ráta által súlyozott mértékben a $\nabla F(x_k)$ gradiens által meghatározott növekedési iránnyal ellentétesen változik:

$$x_{k+1} = x_k - \nabla F(x_k) * \alpha \tag{9}$$

A klasszikus gradiens módszerek esetében az α tanulási ráta értéke minden iterációban konstans, de léteznek olyan megoldások is, melyeknél ez a paraméter iterációként változik, ilyen például az AdaGrad (Adaptive Gradient) [5].

A Q-függvény hangolása során a cél a Q-függvényt leíró Fuzzy-szabálypontok hangolása úgy, hogy az egyes iterációs lépésekben a TD-hiba értékével frissülő Q-függvényt minél kisebb hibával írja le.

A hiba a legkisebb négyzetek módszerével az elvárt kimeneti értékek és a tényleges kimeneti értékek különbségnégyzeteinek az összege, azaz az átlagos négyzetes hiba (Mean Squared Error – MSE) (10):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\big(y_i - F(x_i)\big)^2 \tag{10}$$

ahol y_i az elvárt kimenet leíró a i-edik ($i \in [1, N]$) mintaadat, $F(x_i)$ az F függvény értéke az x_i pontban, N pedig a mintadatok száma.

Jelen esetben hibának a Q-learning TD-hiba értéke tekinthető, amely a jutalom értéke összegezve a diszkontált várható Q-érték és a tényleges Q-érték közötti eltéréssel. A (9) formulában így y_i-nek a $g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a')$, $F(x_i)$-nek pedig a $\tilde{Q}^k(\boldsymbol{s}, a)$ feleltethető meg:

$$TDerror = g(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma * \max_{a' \epsilon U} \tilde{Q}^k(\boldsymbol{s}', a') - \tilde{Q}^k(\boldsymbol{s}, a) \tag{11}$$

A fentiek alapján a gradiens módszerben alkalmazott MSE értéke (melynek a minimalizálása a cél) az alábbi módon határozható meg:

$$MSE = \frac{1}{\hat{m}}\sum_{i=1}^{\hat{m}}(TDerror)^2 \tag{12}$$

A Fuzzy-szabályok antecedens és konzekvens értékeinek hangolása a (9) összefüggés alapján történik, ahol az $F(x_k)$ függvény parciális deriváltja ($\nabla F(x_k)$ gradiens) a láncszabály alkalmazásával a következőképpen határozható meg:

$$\nabla F(x_k) = \frac{\partial MSE(x_k)}{\partial x_k} = \frac{\partial (TDerror)^2}{\partial x_k} = 2 * TDerror * \frac{\partial \tilde{Q}^k(\boldsymbol{s}, a)}{\partial x_k} \tag{13}$$

A (9) összefüggésbe $\nabla F(x_k)$-t a (13) szerint behelyettesítve az x_{k+1} értéke a következő lesz:

$$x_{k+1} = x_k - \left(2 * TDerror * \frac{\partial \tilde{Q}^k(\boldsymbol{s}, a)}{\partial x_k}\right) * \alpha \tag{14}$$

A (14) frissítési szabályt a $\tilde{Q}^k(\boldsymbol{s}, a)$ függvény minden egyes $\boldsymbol{s}, a$ és q dimenziójának parciális deriváltjaira alkalmazva az új $\boldsymbol{s}_{k+1}$ állapot, az új a_{k+1} akció és az új q_{k+1} konzekvens értékek a következő módon számíthatók:

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(\boldsymbol{s}, a)}{\partial \boldsymbol{s}}\right) * \alpha \tag{15}$$

$$a_{k+1} = a_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(\boldsymbol{s}, a)}{\partial a}\right) * \alpha \tag{16}$$

$$q_{k+1} = q_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(\boldsymbol{s}, a)}{\partial q}\right) * \alpha \tag{17}$$

A $\tilde{Q}(\boldsymbol{s}, a)$ függvény parciális derviáltjainak meghatározása, azaz a $\nabla \tilde{Q} = grad(\tilde{Q}) = \left[\frac{\partial \tilde{Q}(\boldsymbol{s},a)}{\partial \boldsymbol{s}}, \frac{\partial \tilde{Q}(\boldsymbol{s},a)}{\partial a}, \frac{\partial \tilde{Q}(\boldsymbol{s},a)}{\partial q}\right]$ gradiens vector számításának megvalósítása numerikusan történik, olyan módon, hogy a $v_j(s_j)$ és a $v(a)$ skálafüggvényeket konstans függvényeknek tekintjük (a probléma egyszerűsítése végett).

A javasolt algoritmus pszeudokódja az alábbi [21]:

Algoritmus: updateRBGradDesc(R, numOfIter, alpha)

Input: rule-base (or a rule), number of iterations, learning rate
Output: tuned rule-base (or a rule)

initialize the weights x as antecedent and consequent of the rules
Repeat
calculate the gradients $\nabla F(x_k)$ with respect to $\boldsymbol{s}, a, q$
update the weights according to $x_{k+1} = x_k - \left(2 * TDerror * \frac{\partial \tilde{Q}^k(\boldsymbol{s},a)}{\partial x_k}\right) * \alpha$
update termination metrics
until the cost stops reducing or number of iterations end
return the tuned rule-base (or a tuned rule)

A bemutatott grandiens módszer alapú szabálybázis-hangolási eljárás következtében előfordulhatnak olyan esetek mikor két szabály állapot-akció pontja közel kerül egymáshoz. Az egymáshoz közel kerülő szabályok nagyon hasonló információt írnak le így célszerű lehet egy olyan szabálybázis-redukálási módszer alkalmazása,

amely a tanulási folyamat során egy javasolt módszer [17] alapján egyetlen szabállyá egyesíti (összevonja) az egymáshoz közel elhelyezkedő Fuzzy-szabályokat, csökkentve ez által a szabálybázis méretét.
A továbbfejlesztett FRIQ-learning módszer, amely alkalmas szakértő által előzetesen definiált tudásbázis FRIQ-learning rendszerbe történő beillesztésére, majd annak hangolására és a tanulási folyamat közbeni redukálására, elnevezése: *Heuristically Accelerated Fuzzy Rule-Interpolation based Q-learning (HFRIQ-learning)* [21].

4. Összefoglalás

Bemutatásra került egy olyan javasolt módszer, amely alkalmas a FRIQ-learning módszer Q-függvényét leíró Fuzzy-szabálybázis hangolására a tanulási folyamat során, abban az esetben, mikor a szakértői által megadott tudásbázis helytelen információt (szakértői szabályokat) tartalmaz. A bemutatott módszer lehetőséget ad a helytelen információt leíró szabályok állapot-akció pontjának (és Q-értékének) optimalizálására a gradiens módszer alkalmazása által, amely következtében a rendszer nem vesz fel számos új szabályt a helytelen szabályok hatásának korrigálása végett, hanem a meglévő szabálypontok hangolását valósítja meg. A szabálybázis-hangolási folyamat során az esetlegesen egymáshoz közel kerülő szabályok a javasolt szabálybázis-redukálási módszer [17] alkalmazása által összevonhatók, a tanulási folyamat során csökkentve a szabálybázis méretét.

További kutatási terv egy olyan szabálybázis-validálási eljárás kidolgozása, amely alkalmas lehet a szakértő által megadott szabályok változásának nyomon követésére a tanulási folyamat során olyan célból, hogy a tanulási folyamat kezdetén definiált szakértői heurisztika, majd a hangolási folyamat végeztével kapott szakértői tudásbázis összehasonlítható legyen, azaz annak helyessége validálására adjon lehetőséget. Az így kialakítandó módszerek jelentősége amellett, hogy egy nyelvi leírási formából kiindulva (például etológiai modell [14], mint a priori tudás) valamilyen rendszert közvetlenül működtető modellként használhatók, megfelelő teljesítménymérték választása és minták megléte esetén a kezdeti szakértői heurisztika validálására is lehetőséget nyújthatnak.

Irodalomjegyzék

[1] 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, https://doi.org/10.1109/FUZZY.1996.553542.

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

[3] Vincze, D., Tóth, A., Niitsuma, M.: Antecedent Redundancy Exploitation in Fuzzy Rule Interpolation-based Reinforcement Learning. *Proc. IEEE/ASME Intl. Conf. on Advanced Intelligent Mechatronics (AIM)*, Boston, USA, 2020, pp. 1316–1321. https://doi.org/10.1109/AIM43001.2020.9158875

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

[5] Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12. 7, 2011.

[6] Rummery, G. A., Niranjan, M.: *On-line Q-learning using connectionist systems*. CUED/F-INFENG/TR 166, Cambridge Univ., UK, 1994.

[7] Haji, Saad Hikmat, Abdulazeez, A. M.: Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18, 4, 2021, pp. 2715–2743.

[8] Kennedy, J., Eberhart, R.: Particle swarm optimization. *Proceedings of ICNN'95-international conference on neural network*, Vol. 4, IEEE, 1995.

[9] Mazyavkina, N., Sviridov, S., Ivanov, S., Burnaev, E.: Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134, 2021, p. 105400, https://doi.org/10.1016/j.cor.2021.105400.

[10] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R.: Heuristically Accelerated Reinforcement Learning: Theoretical and Experimental Results. *Proc of the 20th European Conf. on Artificial Intelligence*, France, 2012, pp 169–174. https://doi.org/10.3233/978-1-61499-098-7-169

[11] Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, USA 1998.

[12] Santra, Santanu, Hsieh, Jun-Wei, Lin, Chi-Fang: Gradient descent effects on differential neural architecture search: A survey. *IEEE Access*, 9, 2021, pp. 89602–89618. https://doi.org/10.1109/ACCESS.2021.3090918

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

[14] Kovács, Sz., Vincze, D., Gácsi, M., Miklósi, Á., Korondi, P.: Ethologically inspired robot behavior implementation. *Proc. 4th International Conference on Human System Interaction (HSI 2011)*, Keio University, Yokohama, Japan, 2011, pp. 64–69. https://doi.org/10.1109/HSI.2011.5937344

[15] Tompa, T., Kovács, Sz.: Applying Expert Heuristic as an a Priori Knowledge for FRIQ-Learning. *Acta Polytechnica Hungarica*, Vol. 17, No. 4, 2020, pp. 27–45.

[16] Tompa, T., Kovács, Sz.: Clustering-based fuzzy knowledgebase reduction in the FRIQ-learning. *Proc. 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI2017)*, Herl'any, Slovakia, 2017, pp. 197–200. https://doi.org/10.1109/SAMI.2017.7880302

[17] Tompa T., Kovács Sz.: Szabálytávolság alapú szabálybázis redukció a szakértői tudásbázissal bővített FRIQ-learning környezetben. *Multidiszciplináris Tudományok*, 12, 1, 2022, pp. 90–102, https://doi.org/10.35925/j.multi.2022.1.8.

[18] Tompa T., Kovács Sz.: Szakértői heurisztika alkalmazása a FRIQ-learning megerősítéses tanulási módszerben. *Multidiszciplináris Tudományok*, 9, 4, 2019, pp. 356–368. https://doi.org/10.35925/j.multi.2019.4.35

[19] Tompa T., Kovács Sz.: Tudásbázis redukció a szakértői szabályrendszerrel bővített FRIQ-learning módszerben. *Multidiszciplináris Tudományok*, 11, 4, 2021, pp. 70–80. https://doi.org/10.35925/j.multi.2021.4.8

[20] Tompa, T., Kovács, S., Vincze, D., Niitsuma, M.: Demonstration of expert knowledge injection in Fuzzy Rule Interpolation based Q-learning. In *2021 IEEE/SICE International Symposium on System Integration* (SII), 2021, January, pp. 843–844, IEEE. https://doi.org/10.1109/IEEECONF49454.2021.9382734

[21] Tompa, T., Kovács, Sz.: Heuristically accelerated FRIQ-learning. *20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2022)*, IEEE, 2022.

[22] Tompa, T., Kovács, Sz.: 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. https://doi.org/10.1109/CarpathianCC.2018.8399677

[23] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al.: Human-level control through deep reinforcement learning. *Nature*, 518, 7540, 2015, p. 529. https://doi.org/10.1038/nature14236

[24] Vincze, D., Kovács, S.: (2009, May). Fuzzy rule interpolation-based Q-learning. In *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, 2009, May, pp. 55–60, IEEE, https://doi.org/10.1109/SACI.2009.5136311.

[25] Vincze, D., Kovács, Sz.: Incremental Rule Base Creation with Fuzzy Rule Interpolation-Based Q-Learning. In: Rudas, I. J. 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

[26] Vincze, D.: Fuzzy rule interpolation and reinforcement learning. *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2017, pp. 173–178, https://doi.org/10.1109/SAMI.2017.7880298.