Multidiszciplináris tudományok, 10. kötet. (2020) 3 sz. pp. 394-402 **https://doi.org/10.35925/j.multi.2020.3.48**

HEURISZTIKUSAN GYORSÍTOTT MEGERŐSÍTÉSES TANULÁSI MÓDSZEREK - ÁTTEKINTÉS

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

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

Absztrakt

A klasszikus megerősítéses tanulási módszerek (Q-learning, SARSA) mindegyike egy megfelelően definiált jutalomfüggvény által, a környezettől kapott visszajelzések ismeretében számos próbálkozással térképezi fel az adott probléma megoldásához vezető utat. A rendszer a tanulási folyamat kezdetén semmilyen tudással nem rendelkezik a probléma megoldásával kapcsolatban, a megoldás tudásbázisát a tanulási fázis során állítja elő, az a célja, hogy iterációról-iterációra feltérképezze azt. Ennek következtében azonban a tanulási folyamat, illetve a probléma megoldása alatt lejátszódó iterációk száma meglehetősen hosszadalmas lehet. Ez a folyamat felgyorsítható lehet abban az esetben, ha áll rendelkezésre részleges információ a megoldásra vonatkozóan és az injektálható a rendszerbe. A heurisztikusan gyorsított megerősítéses tanulási módszerek ember által, valamilyen formában megadott tudást visznek be rendszerbe, amely által a konvergenciasebesség és a megoldás alatt eltelt lépések száma csökkenhető. Jelen cikk célja, hogy áttekintse azon megerősítéses tanulási módszereket, melyek heurisztikusan gyorsítottak, azaz ember által megadott előzetes (a priori) illetve részleges tudásbázis injektálását teszik lehetővé a megerősítéses tanuló rendszerbe.

Kulcsszavak: *megerősítéses tanulás, heurisztikusan gyorsított megerősítéses tanulás, szakértői tudásbázis, Q-learning, fuzzy Q-learning*

Abstract

The conventional reinforcement learning methods (e.g. Q-learning, SARSA) search the solution through trial and error by the properly defined reward function, therefore based on reinforcements given by the environment. The beginning of the learning phase the system does not have any knowledge about the solution, the goal of these methods are to build the knowledgebase during the learning phase. Thus the learning phase can be a long task and the number of iterations which lead to the final solution can be high. In that case the learning process can be speed up, if there are portion of knowledge about the solution and it can be injected into the learning system. The heuristically accelerated reinforcement learning methods incorporate the knowledge defined by human, due to this reason the convergence speed of the system and the number of iterations can be decreased. The main goal of the paper is to give an overview about reinforcement learning methods which heuristically accelerated and give a possibility to inject human defined knowledge into the learning system.

Keywords: *reinforcement learning, heuristically accelerated reinforcement learning, expert knowledgebase, Q-learning, fuzzy Q-learning*

1. Bevezetés

A megerősítéses tanulás (Reinforcement Learning – RL) [22] a mesterséges intelligencia egyre népszerűbb és jobban kutatott tématerülete. Ezen módszerek, algoritmusok működése a környezetből jövő megerősítési információkon (büntetések, vagy jutalmak) alapszik. Az ágens egy megfelelően definiált jutalomfüggvény által, az adott állapotokban végrehajtott akciókra (vagy akciósorozatokra) visszajelzéseket kap a környezettől, majd ezen visszajelzések alapján, kísérletezés útján térképezi fel a megoldáshoz vehető utat. Ezen eljárások nagy előnye abban rejli, hogy a megoldást leíró modell ismerete nélkül, a definiált cél által (jutalomfüggvény formájában) képesek megkeresni a megoldást és létrehozni az azt leíró, működtető tudásbázist. A rendszer tehát megerősítési információk alapján alakítja ki a viselkedést, minden lépésben kap visszajelzést az adott döntés vagy döntések végrehajtását követően, de ezekből arra nem lehet következtetni, hogy ezt mely döntéssorozatának köszönhetően kapta, nincs külső tanár, aki minden esetben adna visszajelzést arról, hogy mi volt a helyes cselekvés. A tanulási módszer alapötlete, hogy a visszajelzéseket ne csak az ágens jelenlegi cselekvésinek kialakítására használják fel, hanem arra is, hogy javítsa a jövőbeli döntésekre irányuló képeséget, azaz a tanulás során, lépésről-lépésre egyre helyesebben oldja meg az adott feladatot. A tanulási folyamat epizodikus, azaz véges hosszúságú időszeletekre (epizódokra) bontott, minden egyes epizód egy kezdeti állapot és egy végállapot között játszódik, jutalmazás az epizódosok végén történik és az egyes epizódok egymástól függetlenek.

A szakirodalomban több megerősítéses tanulási algoritmus is fellelhető, ezek közül a legelterjedtebbek a Q-learning [28], a SARSA [21], illetve ezek különböző változatai. Ezen klasszikus megerősítéses tanulási algoritmusok mindegyike üres tudásbázissal indítja el a tanulási folyamatát, majd iterációról-iterációra bővíti azt, amíg elő nem áll a megoldást leíró végső működtető tudásbázis. A tanulási fázis közben számos epizód játszódhat le, az egyes epizódokon belül pedig tetszőleges, de véges számú iteráció (próbálkozás) lehetséges. A tanulási folyamat közben lejátszódó epizódok száma a konvergencia sebességet határozza meg, tehát, hogy mennyi epizód volt szükséges a rendszer betanításához. A tanulási folyamat ezen rendszerekben hosszadalmas is lehet, nincs semmilyen előzetes információ a probléma megoldására vonatkozóan, a cél ennek megkeresése. Továbbá a konvergencia sebességet a probléma mérete (dimenziószáma) is jelentősen befolyásolja, főleg abban az esetben mikor diszkrét állapot-akció tér van alkalmazva. A problématér mérete egyes szerencsés esetekben hierarchikus állapot-akció tér alkalmazásával is csökkenthető [13][14]. Általánosságban segítene azonban az, ha ezen módszerekbe beágyazható lehetne a működésre vonatkozó előzetes tudásbázis, akkor a tanulási fázis, illetve a megoldás megtalálása alatt eltelt lépések száma nagy mértékben javítható lenne.

Jelen cikk célja összefoglalni azon jelenlegi megerősítéses tanulási módszereket, melyekbe előzetes (a priori) tudásbázis, azaz heurisztika injektálható, javítva ez által a tanulási folyamat hatékonyságát.

2. Klasszikus megerősítéses tanulási algoritmusok

2.1. Q-learning

A Q-learning (Q-tanulás) algoritmus [28], amely eredeti megfogalmazásban diszkrét állapot- és akció tér felbontással rendelkezik - azaz véges számú és diszkrét értékű állapot-akció értékek lehetségesek - a Bellman egyenlet [2] fixpont megoldásait keresi iterációkon keresztül. Az állapot-akció-érték függvény (Q-függvény) frissítési szabálya a következő:

$$Q^{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha * (r_{t+1} + \gamma * \max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (1)$$

Ahol t az adott időpillanat, r_{t+1} a kapott jutalom az $s_t \rightarrow s_{t+1}$ állapotátmenetre, $Q^{new}(s_t, a_t)$ a frissített érték, $Q(s_t, a_t)$ a régi Q-érték (s_t, a_t)-ban, α tanulási ráta (learning rate, $0 < \alpha \leq 1$), γ a diszkontálási tényező (discount factor, $0 < \gamma \leq 1$), $\max_a Q(s_{t+1}, a)$ pedig az a becsült érték, ami az s_{t+1} állapotba vezető feltehetően legjobb a akció végrehajtása mellett érhető el. A $Q(s_t, a_t)$ függvény tehát az adott állapotokban az adott akciók végrehajtása melletti jóság értékeket (Q-érték) adja. Ezen Q-függvény értékei általában egy Q-táblában tárolódnak, amely az összes állapot-akció párra vonatkozó Q-értéket tárolja, majd a tanulási fázis közben a (4) összefüggés alapján frissíti azokat. Minél finomabb az állapot-akció tér felbontása annál nagyobb méretű a Q-tábla, mérete a dimenziók számának és felbontásának növekedésével rohamosan nő.

A Q-learning 'off-policy' algoritmus, amely az (1) formula alapján a Q-értékeket a legjobb akció (mohó akció) alapján frissíti. A mohó akció azt az akciót jelöli, amely végrehajtása mellett az adott állapotban a legnagyobb (legjobb) Q-érték várható. A Q-learning algoritmus tehát mohó akcióválasztási politika alapján frissíti Q-értékeit függetlenül attól, hogy a mohó akció volt-e ténylegesen végrehajtva (mohó politika volt-e alkalmazva):

$$\pi(s) = arg \max_a Q^{\pi}(s, a) \tag{2}$$

A feltérképezés-kiaknázás technika következtében előfordulhat olyan eset mikor nem mindig a feltételezhető legjobb akciót választja a rendszer, hanem adott ε valószínűséggel ($\varepsilon \in [0,1]$) választ véletlen cselekvést is (1-ε értékkel pedig a mohót). Ez az úgynevezett ε–mohó (ε-greedy) politika.
A Q-learning algoritmus pszeudokódja az alábbi [28]:

```
Algorithm parameters: α,γϵ(0,1]
Initialize Q(s,a)=0

Loop (for each episode):
      Initialize s
      Loop for each step of episode:
        Choose a from s using policy derived from Q (e.g., ε-greedy)
            Take action a, observe r, s'
            Q(s,a) ← Q(s,a) + α * (r + γ * max_a Q(s',a) − Q(s,a))
            s ← s'
      until S is terminal
```

2.2. SARSA

A SARSA (State-Action-Reward-State-Action) [21] algoritmus működése hasonló a Q-learning algoritmuséhoz, annak egy módosított verziója, amely egy 'on-policy' módszer. Ebben az esetben a Q-értékek nem mohó politika alapján vannak frissítve, hanem a ténylegesen követett politika, azaz a ténylegesen végrehajtott akció alapján (tehát: Choose a' from s' using policy derived from Q (e. g. , ε − greedy). A SARSA frissítési szabálya a következő:

$$Q(s, a) \leftarrow Q(s, a) + \alpha * (r + \gamma * Q(s', a') - Q(s, a)) \tag{3}$$

2.3. Fuzzy Q-learning

A Fuzzy Q-learning (FQ-learning) [1] [3][9] módszer a diszkrét felbontású Q-learning algoritmus kiterjesztése folytonos állapot-akció térre, fuzzy logika alkalmazásával. A folytonos állapot-akció tér (univerzum) folytonos értékű állapot- és akciódimenziót takar, végtelenszámú lehetséges értéket képviselve az adott dimenzión belül Fuzzy Q-learning esetében a rendszer működtető tudásbázisa nem Q-táblában, hanem fuzzy szabályok formájában van tárolva, a tudásbázis mérete a szabálybázisban lévő szabályok számával egyenértékű. A fuzzy szabályok antecedense az adott állapot és a hozzátartozó akció, konzekvense pedig az ebben a szabálypontban meghatározott Q-érték. Az alkalmazott hagyományos fuzzy következtetés következtében a rendszer működésének feltétele a szabálybázis fedő jellege. Ez azt jeleneti, hogy bármilyen megfigyelés esetén léteznie kell legalább egy olyan szabálynak, amelynek antecedense nagyobb, mint nulla mértékben ($\varepsilon>0$) fedi a megfigyelést minden egyes bemeneti dimenzióban. Tehát elengedhetetlenül fontos, hogy minden létező megfigyelés esetében a rendszer szolgáltasson következtetés. Azonban az állapottér dimenziószámának növekedése a szabálybázis méretének exponenciális növekedéséhez vezet, a rendszer komplexitását növelve ezáltal [16]. Az ezen módszerekben alkalmazott, általában 0-rendű Takagi-Sugeno következtetés, illetve a működtető tudást leíró szabálybázis, mint univerzális függvény approximátor fogható fel, a $\tilde{Q}(s,a)$ függvény közelítő leírásával.

A fuzzy szabályok általános formája a következő:

$$\boldsymbol{If}\ S\ is\ S_i\ \boldsymbol{And}\ A\ is\ A_u\ \boldsymbol{Then}\ \tilde{Q}(s,a) = Q_{i,u} \qquad i \in I, u \in U \tag{4}$$

Ahol, $\tilde{Q}(s,a)$ a folytonos, közelített Q-függvény, $Q_{i,u}$ a szingleton konklúzió, S_i az n-dimenziós állapottér i-edik tagságfüggvénye, A_u az egydimenziós akciótér u-adik tagságfüggvénye.

2.4. Fuzzy szabály-interpoláció alapú Q-learning módszerek

A Fuzzy Q-learning módszerek esetében a rendszer működtető tudását leíró szabálybázis mérete exponenciálisan nő a dimenziószámmal [16]. Ennek következtében a szabályok száma a szabálybázis fedő jellege miatt, bizonyos problémák esetében (dimenziószám függő), bizonyos idő eltelte után (epizódok száma) kezelhetetlen méretűvé válik. A klasszikus 0-rendű Takagi-Sugeno következtetési módszert kicserélve fuzzy szabály-interpolációs (Fuzzy Rule Interpolation - FRI) modellre, a szabálybázis mérete jelentősen csökkenthető, a szabálybázis ritka jellege következtében. A fuzzy szabály-interpolációs módszerek célja, hogy ritka szabálybázisok alkalmazásának esetében is valamilyen módon határozzon meg a rendszer következményt, kimentet.

Az egyik ilyen fuzzy szabály-interpolációs modellt alkalmazó Q-learning módszer a FRIQ-learning [26] [27], de a szakirodalomban több ehhez hasonló módszer is megtalálható [12] [15]. Ezen módszerek általában az alkalmazott fuzzy interpolációs modellben különböznek, kihasználva az adott interpolációs eljárás tulajdonságait. További fuzzy Q-learning és fuzzy interpoláció alapú módszerekről a [27] ad bővebb áttekintést.

3. Heurisztikusan gyorsított megerősítéses tanulási módszerek

A klasszikus megerősítéses tanulási módszerek problémája az esetleges lassú konvergencia sebesség, magas iterációszám [4]. Ennek oka ezen módszerek előnyében keresendő, amely által képesek az állapottér feltérképezésével, próbálkozásokkal megoldást találni egy olyan problémára, melyről kezdetben semmilyen előzetes információ nem állt rendelkezésre. Tehát a rendszer a tanulási fázis kezdetén nem rendelkezik semmilyen előzetes tudásbázissal az adott probléma megoldására vonatkozóan, így az annak méretétől (dimenziószámától) függően több-kevesebb epizód alatt, számos próbálkozással találja meg a

helyes megoldást. A teljesítménymértékek (konvergencia sebesség, megoldáshoz vezető iterációk száma) értéke az állapottér dimenziószámának növekedésével pedig egyre csak növekszik.

Az említett problémák kiküszöbölésére jelenthetnek megoldást azon megerősítéses tanuló rendszerek, amely rendelkeznek valamilyen előzetes (a priori) tudással (heurisztikával) az adott feladat megoldására vonatkozóan. Heurisztika alatt ebben az esetben korábban megszerzett tapasztalat, az adott megerősítéses tanulási feladat megoldására vonatkozó előzetes (és részleges) tudásbázis értendő, amely ember (azaz szakértő) által meghatározott és a rendszer szempontjából külső információ. Fontos megemlíteni, hogy ez az a priori heurisztika általában nem a teljes megoldás leírását jelenti, hiszen abban az esetben a megoldás már ismert, hanem a teljes megoldásnak csak a rendelkezésre álló szeletét, adott részét. Tehát ez az előzetes tudásbázis nem az optimális politikát definiálja és a rendszerhez viszonyítva kívülről származik, nem a tanulási folyamat közben jött létre. Megtalálhatóak olyan módszerek is, melyek a tanulási fázis közben létrejött tudást használják fel újra. Ilyen például a 'Transfer Learning' [25], amely esetében egy korábbi tanulási folyamat során létrejött tudásbázis kerül felhasználásra egy másik, de nagyon hasonló probléma megoldására. Egy másik, már meglévő tudásbázist felhasználó módszer a multiágens rendszer [23], amely estében együttműködő ágensek használják fel egymás tudásbázisait.

Több szerző által is javaslásra kerül a megerősítéses tanulási rendszer valamilyen módon történő előzetes tudásbázissal történő bővítése [5] [8][10][20]. Jelen alfejezet ezen szempontokból tekinti át a témához kapocslódó, publikációkban megtalálható, szakértői információval bővített megerősítéses tanulási módszereket.

3.1. Heurisztikusan gyorsított megerősítéses tanulás

A [5]-ban bevezetett heurisztikusan gyorsított megerősítéses tanulási módszerek az eredeti Q-learning és SARSA [28] algoritmusok módosított változatai. Ezen módszerek rendelkeznek a probléma megoldására vonatkozó részleges tudásbázissal [6], összefoglaló nevük magyar fordítása a „heurisztikusan gyorsított megerősítéses tanulás" (Heuristically Accelerated Reinforcement Learning - HARL). Ebben az esetben egy úgynevezett $H_t(s_t, a_t)$ heurisztikus függvény formájában van definiálva az előzetes heurisztika. Ez a H ($H: S \times A \to R$) függvény egy politika módosítónak tekinthető, azt definiálja, hogy mely s_t állapotban, mely a_t akció végrehajtása preferált az adott t időpillanatban.

A kapcsolat a heurisztikus függvény és az akció-érték függvény között a következő [5]:

$$F_t(s_t, a_t) \bowtie \xi H_t(s_t, a_t)^{\beta} \tag{5}$$

Ahol $F: S \times A \to R$ az értékfüggvény becslése (Q-learning esetében $\tilde{Q}_t(s_t, a_t)$), $H: S \times A \to R$ a heurisztikus függvény, amely az adott a_t akció végrehajtásának preferálását határozza meg s_t-ben, $\bowtie$ függvény, amely a rendezett halmazból állít elő értéket (valós szám), ξ és β pedig a heurisztikus függvény paraméterei, melyek a H függvény rendszerre történő hatását befolyásolják, azaz, hogy H milyen mértékben érvényesüljön.

Heurisztikusan gyorsított megerősítéses tanulási algoritmusok a [5] szerzői által bevezetett HAQL (Heuristically accelerated Q-learning), HA-Q(λ), HA-SARSA(λ) és HA-TD(λ) algoritmusok, melyek az eredeti Q-learning, Q(λ), SARSA(λ) és TD(λ) módszerek heurisztikusan gyorsított változatai [4][5].

A heurisztikusan gyorsított megerősítéses tanulás általános algoritmusának pszeudokódja az alábbi [5]:

```
Produce an arbitrary estimation for the value function.
Define an initial heuristic function H_t(s,a) using an appropriate
method.
Observe the current state s.
Repeat:
      Select an action a by adequately combining the heuristic function and the value
      function.
      Execute action a.
      Receive reinforcement r(s,a) and observe next state s'.
      Update H_t(s,a) using an appropriate method.
      Update value function.
      Update state s ← s'     .
until a stopping criteria is met.
where s = s_t , s'=s_t+1 and a=a_t.
```

3.2. Heurisztika definiálása

Heurisztika definiálása alatt a rendszer szempontjából külső szakértői információ leírásának, illetve az adott formában leírt információ megerősítéses tanulási rendszerbe történő injektálásának módja értendő. A 3.1 alfejezetben bemutatott heurisztikusan gyorsított megerősítéses tanulás esetében a rendszer számára külső információ (heurisztika) egy H heurisztikus függvény formájában definiálható, mint politikamódosító. A H függvény leírását megvalósító módszereket 2 csoportba lehet bontani. Az egyik csoportba azok a módszerek tartoznak melyek korábbi ismereteket alkalmaznak a heurisztika következtetésére, vagy újra felhasználjak egy korábbi feladatban megtanult akcióválasztási politikát („ad hoc" mód). A másik csoportba azon módszerek sorolhatók, melyek a tanulási folyamatból származó információkat használják fel, ilyen lehet az aktuális akcióválasztási politika, az értékfüggvény, állapottér trajektória [5].

Több szerző is javasol más, a heurisztikus függvény leírási módjától eltérő tudásbázis megadási formát, amely alkalmas lehet kezdeti szakértői tudásbázis injektálására a megerősítéses tanuló rendszerbe. Az egyik ilyen lehetséges leírási forma a „GOAL" (Goal-Oriented Agent Language) azaz a célorientált ágens programozási nyelv [11], amely ember számára is olvasható „if then" típusú szabályok által írja le az ágens számára az akcióválasztás módját. Ezen tudásreprezentációs nyelv által különböző névvel ellátott egységekben, kapcsos zárójelek között definiálhatók az adott funkciójú nevesített blokkok. A célállapotok a „goals" nevű blokkban, az előnyben részesített akciók az „actionspec" nevű blokkban, azok várható hatása „pre" és „post" kulcsszóval a blokkon belül, az állapotok a „beliefs" nevű blokkban, az „if then" típusú szabályok pedig a „program" nevű blokkon belül. Ez a GOAL nyelv alkalmas lehet külső (nem a rendszerből származó) információ leírására az ágens viselkedésére vonatkozóan [7]. Az ilyen módon megadott akcióválasztási szabályok a tanulási folyamat során nem változtathatók meg.

A fuzzy szabály alapú megerősítéses tanulási módszerekben (mint például a fuzzy Q-learning vagy a fuzzy szabály-interpoláció alapú Q-learning módszerek) kézenfekvő, hogy a rendszer tudásbázisát leíró fuzzy szabályok formájában lenne célszerű megadni az előzetes szakértői tudásbázist is. Ez a leírási, megadási forma a [18] publikációban is javaslásra kerül.

3.3. Kezdeti Q-érték meghatározása

A diszkrét felbontású Q-learning módszer esetében a Q-táblában tárolt állapot-akció párokra vonatkozó Q-értékek kezdetben (a tanulási folyamat elején) 0 (zéró) értékkel vannak inicializálva. A fuzzy szabályalapú, illetve a fuzzy szabály-interpoláció alapú Q-learning módszerek esetében (például a FRIQ-learning

[26][27]) a kezdeti Q-értékek a szabálybázis kezdeti szabályinak konzekvenseiben jelennek meg zérusként. A fuzzy szabályalapú megerősítéses tanulási rendszerek esetében, az előzetes szakértői tudásbázis szabályaira célszerű lehet valamilyen kezdeti (tanulási fázis előtti), de 0-tól eltérő Q-érték (vagy állapot érték) meghatározása. Erre a szakértői tudásbázis megerősítéses tanuló rendszerbe történő injektálása miatt van szükség, valamint az előzetesen meghatározott Q-érték hatással lehet a rendszer konvergencia sebességére [18][24]. A 0-tól eltérő Q- vagy állapot-érték a szakértői szabály konzekvensében megadott, adott állapotban lévő akció végrehajtásának előnyben részesítését jelzi. Mivel a Q-értékek szakértő által történő meghatározása nehézkes (szinte lehetetlen), így különféle módszerek alkalmazása, kidolgozása szükséges ezen előzetes jóságértékek számításához. Több publikációban is található javaslat illetve módszer kezdeti, azaz a tanulási fázis előtt inicializált Q-érték (vagy állapot érték) meghatározására [17][18][19]. Ez történhet szakértő által leírt, a rendszer szempontjából külső tudásbázis alkalmazása következtében, illetve a tanulási folyamat iterációszámának csökkentése érdekében is.

Egyik lehetséges kezdeti állapot-érték számítási módszer Fuzzy Q-learning alkalmazása esetében az egyes fuzzy univerzumok tagságfüggvényei alapján történhet [12]. Ebben az esetben a szakértői a szabályrendszer által az egyes állapotokban preferált akciók vannak meghatározva, majd ezen állapotokban kerül előzetes állapot-érték kiszámításra. A Q-learning módszer frissítési formulája módosul az előzetesen számított állapot-értékek következtében úgy, hogy a meghatározott állapot érték adott súllyal (β) jelenik meg az összefüggésben [19].

Egy másik hasonló megoldás esetében fuzzy szabályok által, az egyes fuzzy partíciók tagságfüggvényei alapján kerül kezdeti Q-érték meghatározásra a folytonos állapot-akció térrel rendelkező fuzzy Q-learning módszer esetében, majd bemutatásra kerül, hogy a módszer hatékonysága javítható ezáltal [18].

További [17] publikáció a tanulási fázis előtt inicializált Q-értékek (Q_i) hatását vizsgálja a tanulási folyamatra. Egyik esetben egy bináris jutalomfüggvény által határozza meg Q_i értékeket. A bináris jutalomfüggvény által a jutalom mindig végtelen ($r = r_\infty$) kivéve abban az esetben, mikor az aktuális meglátogatott állapot megegyezik a célállapottal, ekkor $r = r_g$. Ha $r_g = r_\infty$ akkor $Q_\infty = r_\infty/(1-\gamma)$, Q_i értéket ezen összefüggés alapján célszerű megválasztani. Folytonos állapottérrel rendelkező Q-learning módszer esetében javaslásra kerül folytonos jutalomfüggvény alkalmazása, amely következtében a kezdeti Q-értékek például Gauss eloszlási függvény alkalmazásával kerül inicializálásra. Egy másik lehetséges eset mikor ugyanazon értékekkel kerül inicializálásra Q_i, ekkor $Q_i = \beta/(1-\gamma)$, ahol β konstans érték [17].

4. Összefoglalás

A cikkben áttekintésre kerültek azon megerősítéses tanulási módszerek, melyek szakértői tudásbázis alkalmazására adnak lehetőséget, illetve ezek mellett bemutatásra kerültek a klasszikus megerősítéses tanulási eljárások is. Az áttekintett módszerek által, a rendszer szempontjából külső információként, előzetes szakértői tudásbázis formájában megadott információ rendszerbe történő injektálásával lehetőség nyílik a megerősítéses tanuló rendszerek konvergencia sebességének javítására, illetve a megoldáshoz vezető lépések számának csökkentésére. Bemutatásra kerültek továbbá az előzetes szakértői tudásbázis leírási módjának lehetséges formái, illetve az ehhez szükséges kezdeti Q-érték (és állapot-érték) inicializálási módszerek.

5. Köszönetnyilvánítás

A cikkben ismertetett kutató munka az EFOP-3.6.1-16-2016-00011 jelű „Fiatalodó és Megújuló Egyetem – Innovatív Tudásváros – a Miskolci Egyetem intelligens szakosodást szolgáló intézményi fejlesztése” projekt részeként – a Széchenyi 2020 keretében – az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

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] Bellman, R. E.: *Dynamic programming*, Princeton University Press, Princeton, NJ 1957

[3] 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.

[4] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R.: *Heuristically accelerated reinforcement learning: Theoretical and experimental results*, ECAI. 2012.

[5] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R.: *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**

[6] Bianchi, R. A. C., Ribeiro, C. H. C., Costa, A. H. R.: *Heuristically accelerated Q–learning: a new approach to speed up Reinforcement Learning*, Brazilian Symposium on Artificial Intelligence, Springer, Berlin, Heidelberg, 2004. **https://doi.org/10.1007/978-3-540-28645-5_25**

[7] Broekens, J., Hindriks, K., Pascal, W.: *Reinforcement learning as heuristic for action-rule preferences*, International Workshop on Programming Multi-Agent Systems, Springer Berlin Heidelberg, 2010.

[8] Brys, T.: *Reinforcement Learning with Heuristic Information*, Diss. PhD thesis, PhD thesis, Vrije Universitet Brussel, 2016.

[9] Glorennec, P. Y., Jouffe, L.: (1997, July). *Fuzzy Q-learning*, In Proceedings of 6th International Fuzzy Systems Conference (Vol. 2, pp. 659-662). IEEE.

[10] Hailu, G., Sommer, G.: *Embedding knowledge in reinforcement learning*, International Conference on Artificial Neural Networks. Springer, London, 1998. **https://doi.org/10.1007/978-1-4471-1599-1_178**

[11] Hindriks, K. V., De Boer, F. S., Van Der Hoek, W., Meyer, J. J. C.: (2000, July). *Agent programming with declarative goals*, In International Workshop on Agent Theories, Architectures, and Languages (pp. 228-243), Springer, Berlin, Heidelberg. **https://doi.org/10.1007/3-540-44631-1_16**

[12] 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.

[13] Jeni, L. A., Istenes, Z., Korondi, P., Hasimoto, H.: *Hierarchial reinforcement learning for robot navigation using the intelligent space concept*, Proceedings of the 11th IEEE International Conference on Intelligent Engineering Systems, IEEE Press, 2007, pp. 149-153. **https://doi.org/10.1109/INES.2007.4283689**

[14] Jeni, L. A., Istenes, Z., Korondi, P., Hashimoto, H.: *Mobile agent control in intelligent space using reinforcement learning*, Proceedings of the 7th IEEE International Symposium of Hungarian Researchers on Computational Intelligence, HUCI 2006, Budapest, Hungary, 2006, pp. 201-210.

[15] Kim, M. S., Hong, G. G., Lee, J. J.: (1999, October). *Online fuzzy Q-learning with extended rule and interpolation technique*, In 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, pp. 757-762). IEEE.

[16] Kóczy, L. T., Hirota, K.: *Size reduction by interpolation in fuzzy rule bases*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 27, 14 - 25, 1997. **https://doi.org/10.1109/3477.552182**

[17] Matignon, L., Laurent, G. J., Le Fort-Piat, N.: *Reward function and initial values: better choices for accelerated goal-directed reinforcement learning*, International Conference on Artificial Neural Networks, Springer, Berlin, Heidelberg, 2006. **https://doi.org/10.1007/11840817_87**

[18] Oh, Chi-Hyon, Nakashima, T., Ishibuchi, H.: *Initialization of Q-values by fuzzy rules for accelerating Q-learning*, 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227). Vol. 3. IEEE, 1998.

[19] Pourhassan, M., Mozayani, N.: *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**

[20] Ribeiro, Carlos HC.: *Embedding a priori knowledge in reinforcement learning*, Journal of Intelligent and Robotic Systems 21.1 (1998): 51-71. **https://doi.org/10.1023/A:1007968115863**

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

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

[23] Tan, M.: *Multi-agent reinforcement learning: Independent vs. cooperative agents*, Proceedings of the tenth International Conference on Machine Learning. 1993. **https://doi.org/10.1016/B978-1-55860-307-3.50049-6**

[24] Tompa, T., Szilveszter Kovács, Sz.: *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**

[25] Torrey, L., Shavlik, J.: *Transfer learning, Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010. 242-264. **https://doi.org/10.4018/978-1-60566-766-9.ch011**

[26] 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**

[27] 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**

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