



Q-ÉRTÉK INICIALIZÁLÁS A HFRIQ-LEARNING RENDSZERBEN

TOMPA TAMÁS

Miskolci Egyetem

Informatikai Intézet

Általános Informatikai Intézeti Tanszék

tamas.tompa1@uni-miskolc.hu

KOVÁCS SZILVESZTER

Miskolci Egyetem

Informatikai Intézet

Általános Informatikai Intézeti Tanszék

szilveszter.kovacs@uni-miskolc.hu

Absztrakt.

A Heurisztikusan Gyorsított Fuzzy Szabály-Interpoláció alapú Q-tanulás (HFRIQ-learning) módszer célja a szakértő által megadott tudásbázis injektálása a tanulási folyamatba (a tanulási fázis előtt), majd ezen kezdeti tudásbázis hangolása, pontosítása a tanulási folyamat során. A szakértő által megadható szabályok „ha-akkor” formátumúak, ahol a „ha” rész az állapot, az „akkor” rész pedig az ebben az állapotban preferált akció. A szakértői szabályrendszer HFRIQ-learning rendszerbe történő injektáláshoz minden egyes szakértői szabályt „állapot-akció-Q-érték” formátumúra szükséges konvertálni, amely következtében szükséges a szabályok „Q-érték” részének meghatározása. A cikk célja egy kezdeti Q-érték inicializálási módszer bemutatása, amely a szakértő által definiált minden egyes szabályra kezdeti Q-értéket határoz meg, amely következtében a szakértői szabályrendszer „állapot-akció-Q-érték” formátumban injektálható a tanulási folyamatba.

Kulcsszavak: megerősítéses tanulás, Q-learning, Fuzzy Q-learning, Q-érték inicializálás, szakértői tudásbázis

1. Bevezetés

A megerősítéses tanulás (Reinforcement Learning, RL) [5] egy olyan gépi tanulási paradigma, amelyben az ágens a környezetével való interakciók és az interakciók során kapott visszajelzések (megerősítések) alapján térképezi fel az adott probléma megoldását. Ezen módszereket trial-and-error (próbálkozás) típusú módszerek is nevezik, mert akciók (cselekvések) próbálgatása során kapott megerősítési értékek alapján olyan jövőbeli akciók választására törekszik a módszer, melyek maximalizálják a szerezhető jutalmakat.

Ilyen megerősítéses tanulási algoritmus a Q-tanulás (Q-learning) [16], illetve a SARSA [4] is, amelyek diszkrét állapot- és akció tér felbontással rendelkeznek

(azaz véges számú és diszkrét értékű állapot-akció érték lehetséges). Ezen algoritmusok közös jellemzője, hogy a Bellman egyenlet [1] fixpont megoldásait keresik számos iteráción keresztül és az adott állapotokban választható akciók hasznosságát egy Q-értékkel (hasznosság értékkel) jellemzik. A Q-érték az adott állapotban az adott akció végrehajtása melletti jóság értékeket határozza meg, tehát azt, hogy az adott akció végrehajtása mennyire eredményes („mennyit ér”) az adott állapotban, figyelembe véve a jövőbeli várható jutalmakat is. Minden egyes lehetséges állapot-akció párra ez egy $Q(s_t, a_t)$ Q-függvényt eredményez, amely értékei általában (diszkrét állapot-akció tér esetében) egy Q-táblában tárolódnak, majd a tanulási fázis során pedig iteratív módon kerülnek frissítésre. A Q-értékek tehát ezen algoritmusok alapvető elemei, a tanulási folyamat során létrejött tudásbázist reprezentálják.

A Heurisztikusan Gyorsított Fuzzy Szabály-Interpoláció alapú Q-tanulás (HFRIQ-learning) [8] módszer esetében a rendszer tudásbázisát egy ritka (interpolált) fuzzy szabálybázis írja le, ahol a fuzzy szabályok konzekvens része a Q-érték, antecedense pedig az állapot-akció. Az állapot-akció formátumú szakértői szabályok tanulási folyamatba történő beillesztéséhez szükség van azok Q-értéknek (nem szakértő által történő) meghatározására.

A cikk célja egy olyan Q-érték inicializálási módszer bemutatása, amely alkalmas a szakértő által definiált állapot-akció formátumú fuzzy szabályokra Q-értéket meghatározni, amely következtében azok már adaptálhatók a HFRIQ-learning rendszer tanulási folyamatába.

2. A HFRIQ-learning megerősítéses tanulási módszer

A heurisztikusan gyorsított FRIQ-learning (Heuristically Accelerated Fuzzy Rule-Interpolation based Q-learning – HFRIQ-learning) [8] a FRIQ-learning (Fuzzy Rule-Interpolation based Q-learning) [13][15] továbbfejlesztett változata, amely lehetőséget biztosít külső szakértői tudásbázis integrálására [6], valamint annak finomhangolására [10]. A módszer a Q-függvény reprezentálásához a „FIVE” (Fuzzy Rule Interpolation based on Vague Environment) [2][3] szabály-interpolációs algoritmust alkalmazza, amely következtében módszer állapot-akció tere folytonos.

A rendszer tudásbázisának egyetlen r_i ($i \in [1, m]$) szabályának formátuma az m méretű R szabálybázisban a következő [9][13]:

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

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 $\mathcal{S}$ állapottérben, $s \in \mathcal{S}$ az n -dimenziós állapot megfigyelés, s_j a j -edik dimenziója az s állapot megfigyelésnek, A^i az i -edik szabály egydimenziós akció univerzumának (U) fuzzy halmaza, $a \in U$ az akció, $\tilde{Q}(s, a)$ a FIVE FRI [2][3] által becsült Q-függvény, q^i pedig az i -edik szabály konzekvensé (Q-értéke).

A szakértői tudásbázis (R_{expert}) formátuma hasonló az (1) formula által meghatározott szabályokéhoz, azzal az eltéréssel, hogy a szakértői szabályok ($\hat{r}$) antecedense az állapot, konzekvensé pedig az ebben az állapotban preferált akció [6]:

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

ahol $\hat{r}_i$ az i -edik ($i \in [1, \hat{m}]$) szakértői szabály az R_{expert} szabálybázisban, $\hat{S}_n^i = [\hat{S}_1^i, \hat{S}_2^i, \dots, \hat{S}_n^i]$ az i -edik szakértői szabály n -dimenziós állapot megfigyelése, $\hat{A}^i$ az ehhez az $\hat{S}_n^i$ állapot megfigyeléshez tartozó akció, i ($i \in [1, \hat{m}]$) pedig a szabály indexe az $\hat{m}$ méretű szakértői szabályrendszerben.

Annak következtében, hogy a szakértői szabályrendszer injektálható legyen a rendszerbe formátumának átalakítása szükséges „állapot-akció-Q-érték” formátumra. Ezután az átalakított szakértői szabályok antecedense az állapot-akció, konzekvense pedig egy becsült $\tilde{Q}_{init}$ érték lesz. A cikk célja ezen $\tilde{Q}_{init}$ érték számítási módjának bemutatása, amely a későbbi 3. fejezetben kerül részletesen bemutatásra.

A HFRIQ-learning tanulási folyamata egy kezdeti fuzzy szabálybázissal indul, amely a szakértői szabályok és az állapottér sarokpontjaiban elhelyezkedő, 0 konzekvens értékű szabályok kombinációjával jön létre. A szakértői szabályok és a sarokponti szabályok között lévő esetleges ellentmondásokat a rendszer az ellentmondó szabályok összevonással oldja fel. A tanulás során új szabályok akkor kerülnek beillesztésre a szabálybázisba, ha a Q-érték változása meghalad egy előre meghatározott küszöbértéket, és nincs az aktuális megfigyeléshez közeli, már létező szabály. Ha a Q-érték változása (frissítése) kicsi, akkor a meglévő szabályok konzekvensai frissülnek a következő összefüggés szerint [13][15]:

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

$$\Delta \tilde{Q}^{k+1}(s, a) = \alpha * (g(s, a, s') + \gamma * \max_{a' \in U} \tilde{Q}^k(s', a') - \tilde{Q}^k(s, a)) \quad (4)$$

ahol $\gamma \in [0, 1]$ a leszámítási 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 s -ben végrehajtott akció. Az új megfigyelt állapot s' , $g(s, a, s')$ a megfigyelt jutalom az $s \rightarrow s'$ állapot átmenetre, $\tilde{Q}^k$ és $\tilde{Q}^{k+1}$ pedig a k -edik és a $(k + 1)$ -edik iteráció FIVE FRI módszer által becsült Q-értéke [13][15]:

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

ahol q^i az i -edik ($i \in [1, m]$) szabály konklúziója, (s, a) a megfigyelés, δ_v^i a skálázott távolság [3] az (s, a) megfigyelés és az i -edik szabály (s^i, a^i) antecedense között, λ a Shepard paraméter, m pedig a szabályok száma.

Ha a Q-érték változása a frissítés során kicsi és van már létező szabály a megfigyelés közelében [8], akkor egy gradiens módszer alapú hangolási eljárás a megfigyeléshez legközelebb lévő szabálypont antecedensét és konzekvensét fogja hangolni [10] a következő összefüggések szerint [8][10]:

$$s_{k+1} = s_k - \left(2 * TDerror * \frac{\partial \tilde{Q}(s, a)}{\partial s} \right) * \alpha \quad (6)$$

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

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

ahol a $s_{k+1}, a_{k+1}, q_{k+1}$ a gradiens-módszer által meghatározott új állapot, akció és Q-értékek, s_k, a_k, q_k a régi állapot, akció és Q-értékek, α a gradiens-módszer tanulási rátája, $\frac{\partial \tilde{Q}(s, a)}{\partial s}, \frac{\partial \tilde{Q}(s, a)}{\partial a}, \frac{\partial \tilde{Q}(s, a)}{\partial q}$ a Q-függvény állapot, akció és Q-érték szerinti parciális deriváltjai, a $TDerror$ értéke pedig a következő [8][10]:

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

A tanulási folyamat során alkalmazott szabályhangolás következtében előfordulhat, hogy egyes szabálypontok közel kerülnek egymáshoz. Ilyen esetekben, mivel ezek a szabálypontok nagyon hasonló információt írnak le, a rendszer ezeket egyetlen szabállyá vonja össze a tanulási folyamat során, csökkentve ezáltal a szabálybázis méretét [8][10]. További, a tanulási fázis után opcionálisan alkalmazható szabálybázis-csökkentő eljárásokat a [7][11][12] és [14] források mutatnak be.

A rendszer tanulási fázisa akkor tekinthető befejezettnek, ha már nem jön létre új szabály, a Q-értékek változása és a szabálypontok vándorlása minimális, illetve nem történik már újabb szabályösszevonás (nincsennek egymáshoz közeli szabályok).

3. Q-érték inicializációs módszer a szakértői szabályokra

Az előzetes tudásbázis a szakértő által szabályrendszer formájában kerül leírásra a (2) formula által meghatározott módon. A szabályrendszerben az antecedens a többdimenziós állapot univerzumot, a konzekvens pedig az akció dimenziót jelöli, azaz ezen szabályok definiálásukkor nem rendelkeznek Q-értékkel. A HFRIQ-learning módszer szabályrendszere a (1) formula szerint állapot-akció-Q-érték formátumú, ahol az antecedens a többdimenziós állapot és az akció, a konzekvens pedig a Q-érték.

Annak érdekében, hogy a szakértői szabályok a HFRIQ-learning szabályrendszerébe illeszthetők legyenek, meg kell határozni a szakértői szabályok kezdeti Q-érték konzekvensét. A szakértői szabályok akció konzekvenssei az antecedens oldalra kerülnek, majd az új konzekvensük pedig ezen Q-érték lesz. Ezt a folyamatot még a tanulási fázis megkezdése előtt kell megvalósítani. A kezdeti Q-érték meghatározási módszer célja tehát még a tanulási fázis előtt, a szakértői szabályrendszer minden egyes szabályára becsült Q-érték ($\tilde{Q}_{init}$) inicializálása, azaz a kezdeti Q-függvény meghatározásra. Ez a szakértői heurisztikából létrehozott kezdeti Q-függvény lesz hangolva a tanulási folyamat során.

Feltételezve, hogy a szakértő által megadott szabályok megkérdőjelezhetetlenül helyesek, az azokra meghatározott Q-értékeknek relatívan magasnak (és lehetőleg 0-tól eltérőnek) kell lenniük. A relatívan magas érték egy kezdeti becslés, amely a környezet által maximálisan adható jutalom (g_{max}) ismeretében következők alapján határozható meg:

Jelöljük a szakértői szabályok kezdeti becsült Q-értékét $\tilde{Q}_{init}$ -el. Ez a kezdeti Q-

érték becslés pedig legyen a maximálisan szerezhető Q-érték ($\tilde{Q}_{max}$) érték η -ed ($\eta \in [0,1]$) része:

$$\tilde{Q}_{init} = \eta * \tilde{Q}_{max} \quad (10)$$

Ahol η azt határozza meg, hogy a maximálisan szerezhető, becsült Q-érték mekkora része legyen a szakértői szabályok tényleges konzekvense értéke (Q-értéke). A $\tilde{Q}_{max}$ értékét a következőképpen lehet kifejezni a (4) frissítési formula alapján:

$$\tilde{Q}_{max} = \lim_{k \rightarrow \infty} \tilde{Q}^{k+1}(s^*, a^*) \quad (11)$$

$$\tilde{Q}_{max} = \lim_{k \rightarrow \infty} (\tilde{Q}^k(s^*, a^*) + \alpha * g(s^*, a^*, s^*) + \gamma * \tilde{Q}^k(s^*, a^*) - \tilde{Q}^k(s^*, a^*))$$

Ahol legyen $\tilde{Q}^k(s^*, a^*) = \max_{a' \in U} \tilde{Q}^k(s^*, a')$ és $g(s^*, a^*, s^*) = \max_{s \in S, a \in U} g(s, a, s') = g_{max}$. α a tanulási ráta, γ a diszkontálási tényező, $g(s, a, s')$ a jutalom értéke az $s \rightarrow s'$ állapotátmenetre, $\tilde{Q}^k$ a k -adik, $\tilde{Q}^{k+1}$ pedig a $k+1$ -edik iteráció becsült konklúziója. A g_{max} paraméter, amely szintén a szakértő által meghatározott és a környezet által adható maximális megerősítés értékét jelöli. Ez a (11) összefüggés azt mutatja meg, hogy miközben az iterációk száma (k) közelít a végtelenhez a Q-érték hova konvergál.

A (11) összefüggést tovább fejte a következő formulához jutunk:

$$\begin{aligned} \tilde{Q}_{max} &= \lim_{k \rightarrow \infty} (\tilde{Q}^k(s^*, a^*) + \alpha * (g_{max} + (\gamma - 1) * \tilde{Q}^k(s^*, a^*))) = \\ &= \tilde{Q}^k(s, a) + \alpha * g(s, a, s') + \alpha * (\gamma - 1) * \tilde{Q}^k(s', a') = \\ &= \frac{\alpha * g_{max}}{-\alpha * (\gamma - 1)} = \frac{g_{max}}{1 - \gamma} \end{aligned} \quad (12)$$

A fenti levezetés alapján a szakértői szabályok kezdeti becsült Q-értéke ($\tilde{Q}_{init}$) a környezet által adható maximális megerősítés (g_{max}) és a 1-diszkontálási faktor (γ) hányadosa:

$$\tilde{Q}_{init} = \eta * \frac{g_{max}}{1 - \gamma}, \text{ ahol } \gamma < 1 \quad (13)$$

Ahol $\tilde{Q}_{init}$ a becsült, maximálisan elérhető Q-érték a környezet által maximálisan adható jutalom g_{max} ismeretében, $\eta \in [0,1]$ a $\tilde{Q}_{init}$ skála faktora, amely azt határozza meg, hogy a számított $\tilde{Q}_{init}$ érték mekkora része (százaléka) kerüljön figyelembevételre, γ pedig a diszkontálási tényező. Az így kiszámított $\tilde{Q}_{init}$ érték a szakértői szabályrendszer minden egyes szabályára, azaz annak konzekvens részére egyforma értékű.

A szabályrendszer formája az előzetes Q-érték meghatározási módszer alkalmazása után az (1) összefüggés mintájára a következőképpen módosul:

$$\hat{r}_i: \text{If } s_1 \text{ is } \hat{S}_1^i \text{ And } s_2 \text{ is } \hat{S}_2^i \text{ And ... And } s_n \text{ is } \hat{S}_n^i \text{ And } a = \hat{A}^i \text{ Then } \tilde{Q}(s, a) = \tilde{Q}_{init} \quad (14)$$

4. Összefoglalás

A cikk egy olyan Q-érték inicializálási módszert mutatott be, amely lehetővé teszi a HFRIQ-learning rendszerben megadott szakértői szabályokra történő kezdeti Q-értékek meghatározását. Az eljárás célja, hogy az állapot-akció formátumú szakértői szabályokat a módszer hatékonyan integrálja a tanulási folyamatba. Ennek során a szakértői szabályok állapot-akció-Q-érték formára kerülnek átalakításra, majd a Q-érték rész kezdeti értékének meghatározása a javasolt módszer által történik, amely a szakértő által megadott maximális megerősítési értéken alapul. A bemutatott módszer hozzájárul a szakértői tudásbázis HFRIQ-rendszer tanulási folyamatába történő injektálásához és a szakértői tudás hatékony felhasználásához.

Irodalomjegyzék

- [1] Bellman, R. E.: Dynamic Programming. Princeton University Press, Princeton, NJ 1957
- [2] Kovács, Sz., Kóczy, L. T.: The use of the concept of vague environment in approximate fuzzy reasoning. Fuzzy Set Theory and Applications, Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Bratislava, Slovak Republic, vol.12, 1997, pp. 169-181.
- [3] Kovács, Szilveszter. "Extending the fuzzy rule interpolation" five" by fuzzy observation." *Computational Intelligence, Theory and Applications: International Conference 9th Fuzzy Days in Dortmund, Germany, Sept. 18–20, 2006 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. https://doi.org/10.1007/3-540-34783-6_48
- [4] Rummery, G. A., Niranjan, M.: On-line Q-learning using connectionist systems. CUED/F-INFENG/TR 166, Cambridge University, UK., 1994
- [5] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018. <https://doi.org/10.1017/s0263574799211174>
- [6] Tomba, Tamás, and Szilveszter Kovács. "Applying Expert Heuristic as an a Priori Knowledge for FRIQ-Learning." *Acta Polytechnica Hungarica* 17.4 (2020). <https://doi.org/10.12700/aph.17.4.2020.4.2>
- [7] Tomba, Tamás, and Szilveszter Kovács. "Clustering-based fuzzy knowledgebase reduction in the FRIQ-learning." 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE, 2017. <https://doi.org/10.1109/sami.2017.7880302>
- [8] Tomba, Tamás, and Szilveszter Kovács. "Heuristically accelerated FRIQ-learning." 20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2022). IEEE, 2022. <https://doi.org/10.1109/iccc57093.2023.10178919>
- [9] Tomba, Tamás, and Szilveszter Kovács. "Integrating Expert Knowledge into Fuzzy Reinforcement Learning." 2024 IEEE 18th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE, 2024. <https://doi.org/10.1109/saci60582.2024.10619888>
- [10] Tomba, Tamás, and Szilveszter Kovács. "Knowledge Base Optimization of the HFRIQ-Learning." *Acta Polytechnica Hungarica* 21.10 (2024). <https://doi.org/10.12700/aph.21.10.2024.10.6>
- [11] Tomba, Tamás, and Szilveszter Kovács. "Tudásbázis redukálás a heurisztikusan gyorsított FRIQ-learning rendszerben." *Production Systems and Information Engineering* 11.2 (2023): 1-12. <https://doi.org/10.32968/psaie.2022.4.4>
- [12] Vincze, Dávid, Alex Tóth, and Mihoko Niitsuma. "Antecedent redundancy exploitation in fuzzy rule interpolation-based reinforcement learning." 2020

- IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM).
<https://doi.org/10.1109/aim43001.2020.9158875>
- [13] Vincze, Dávid, and Szilveszter Kovács. "Fuzzy rule interpolation-based Q-learning." 2009 5th International Symposium on Applied Computational Intelligence and Informatics. IEEE, 2009. <https://doi.org/10.1109/saci.2009.5136311>
- [14] Vincze, Dávid, and Szilveszter Kovács. "Rule-base reduction in Fuzzy Rule Interpolation-based Q-learning." Recent Innovations in Mechatronics 2.1-2. (2015): 1-6. <https://doi.org/10.17667/riim.2015.1-2/10>.
- [15] Vincze, Dávid.: "Fuzzy Rule Interpolation-based Q-learning." PhD dissertation, 2013.
- [16] Watkins, C. J. C. H., Dayan, P.: Q-learning, in Machine Learning, Vol. 8 (3/4), 1992., pp. 279-292.

EARLY ACCESS