Multidiszciplináris tudományok, 4. kötet. (2014) 1. sz. pp. 135-146.

AUTOMATIKUS TUDÁSKINYERÉS FUZZY SZABÁLY INTERPOLÁCIÓ ALAPÚ Q-TANULÁSSAL

Vincze Dávid
egyetemi adjunktus, Miskolci Egyetem, Informatikai Intézet,
Általános Informatikai Intézeti Tanszék
3515 Miskolc, Miskolc-Egyetemváros, e-mail: vincze.david@uni-miskolc.hu

Összefoglalás

Jelen cikk egy olyan új eljárást mutat be, amely képes automatikus tudáskinyerésre olyan esetekben, amikor egy rendszer egzakt működése nem ismert, illetve nem áll rendelkezésre adott bemeneti halmazhoz tartozó kimeneti halmaz, tehát mintaadatok alapján nem lehet rendszerműködés generálást végezni. Az eljárás alapja egy már korábban kifejlesztett megerősítéses tanulási módszer, illetve annak egy inkrementális szabálybázis konstrukciós kiegészítése. Az utóbbi módszerekkel eredményül kapott szabálybázisokat felhasználva történik a végleges tudáskinyerés különböző, újonnan kifejlesztett, dekrementális szabálybázis redukciós eljárásokkal. A cikk első része egy áttekintést ad az említett megerősítéses tanulási módszerről, a második része bemutatja az új algoritmusokat, melyek eredményeit egy alkalmazás példán keresztül szemlélteti.

Kulcsszavak: *Q-tanulás, fuzzy szabály interpoláció, szabálybázis redukció, tudáskinyerés*

Abstract

This paper introduces a novel method which is suitable for automatic knowledge extraction in cases when the exact operation of the system is unknown and there is no data available regarding the output set of a given input set, so system generation cannot be done based on sample data sets. The new technique is based on a previously developed reinforcement learning method and its extension, which is able to construct a rule-base incrementally from scratch. The final knowledge extraction is achieved using newly developed decremental rule-base reduction strategies, which make use of the resulting rule-base of the former method. The first part of the paper gives a brief overview of the mentioned reinforcement learning method. The second part introduces the new algorithms alongside with an application example of this new method.

Keywords: *Q-learning, fuzzy rule interpolation, rule-base reduction, knowledge extraction*

1. Bevezetés

Olyan rendszerekben, ahol maga a működés egzakt folyamata nem ismert, különböző tudáskinyerési módszerekkel feltárható a rendszer vagy annak egy részének a működtető tudásbázisa. A tudáskinyerés egyik módja lehet megerősítéses tanulási módszerek (Reinforcement Learning – RL) alkalmazása megfelelően definiált jutalomfüggvényekkel egyetemben. Ilyen esetekben megvalósítható a tudáskinyerés automatizálása is, hiszen a

jutalomfüggvény alapján meghatározható, hogy egy adott aktuális tudásbázis helyesen képes-e megoldani a feladatot, vagy sem.

A megerősítéses tanulás előnye, hogy a megoldandó problémánál nem a megoldás egzakt lépésenkénti mikéntjét, hanem az elérendő végcélt definiálhatjuk jutalomfüggvény formájában. Így maga a probléma megoldása a környezettől kapott visszajelzésekben rejlik (jutalomfüggvény). Ezen visszajelzések (jutalmak / büntetések) felhasználásával a rendszer képes arra, hogy felderítse azokat a beavatkozásokat, amelyek a legjobbnak bizonyulnak egy-egy adott állapotban. Az egyik leggyakrabban alkalmazott megerősítéses tanulási módszer a Q-tanulás (Q-learning), amely eredeti megfogalmazásában csak diszkrét felbontású terekben alkalmazható, fuzzy következtetés bevezetésével azonban kiterjeszthető folytonos terekre is (Fuzzy Q-tanulás). Lényegesen csökkenthető a fuzzy modell komplexitása a ritka szabálybázis alkalmazását lehetővé tevő fuzzy szabály interpolációs módszerek bevezetésével. Ez utóbbit alkalmazza az FRIQ-tanulás is, ami egy fuzzy szabály interpoláció alapú Fuzzy Q-tanulási módszer.

A következőekben ezen módszerek rövid áttekintése olvasható, továbbá egy olyan új eljárás bemutatása, amely az FRIQ-tanulás eddig elért eredményeit hasznosítja és bővíti ki, alkalmassá téve automatikus tudásbázis kinyerésre a kifejlesztett dekrementális szabálybázis redukciós módszerek segítségével.

2. Fuzzy szabály interpoláció alapú Q-tanulás (FRIQ-tanulás)

A megerősítéses tanulási módszerek általában azokban a helyzetekben használhatóak eredményesen, amikor az adott feladat megoldása kinyerhető a környezet visszajelzéseiből, azaz jutalmakból, amelyek lehetnek pozitívak vagy negatívak (ebben az esetben szokás büntetésnek nevezni). Természetesen a jutalmakat megadó függvényt az adott probléma elvárt megoldásával összhangban kell meghatározni. A környezet által visszaadott jutalmak alapján, a megerősítéses tanulási módszerek képesek feltérképezni azokat az akciókat, melyek az egyes rendszerállapotokból a kívánt megoldások felé vezetnek.

Ezek a módszerek iterációról iterációra haladva próbálgatják végig a lehetséges akciókat és a kapott jutalom alapján becslik meg a legnagyobb jutalommal kecsegtető lépéseket. A megerősítéses tanulási módszerek koncepcionálisan a dinamikus programozás területéről [3] származnak, közös céljuk az állapot-akció-érték függvény (adott állapotban megadja a választható akciók minőségi értékét) meghatározása, vagyis egy optimális stratégia megtalálása [12].

Az optimális stratégia becsléséhez az akció-érték függvényt kell közelíteni, ami összetett feladat, mivel mind a lehetséges állapotok, mind a lehetséges akciók száma rendkívül magas lehet. Általánosságban elmondható, hogy a megerősítéses tanulási módszerek csak viszonylag kis állapot-, illetve akciótérben alkalmazhatóak sikeresen (nagyobb állapottérben belátható időn belül vezetnének eredményhez).

A megerősítéses tanulási módszerek egyike a Q-tanulás (Q-learning) [15], melynek célja a Bellman egyenlet [3] iterációkon keresztüli megoldása. Az algoritmus eredetileg diszkrét terek alkalmazásához lett meghatározva, de fuzzy modellek alkalmazásával a diszkrét Q-tanulás átalakítható úgy, hogy folytonos állapot és akció tereken is alkalmazható legyen. Több megoldás is létezik a Q-tanulás folytonos térre való kiterjesztésére fuzzy következtetési rendszerek alkalmazásával [1], [4], [5], [6]. A legegyszerűbb Fuzzy Q-tanulás (FQ-tanulás) a nulladrendű Takagi-Sugeno fuzzy következtető modellt alkalmazza, ahol az akció-érték függvény jellemzésére folytonos állapot-akció (s, a) térben a

nulladrendű Takagi-Sugeno fuzzy következtető modelles közelítés $\tilde{Q}(s,a)$ a következő formában adódik:

$$\textbf{HA } s = S_i \textbf{ ÉS } a = A_u \textbf{ AKKOR } \tilde{Q}(s,a) = Q_{i,u} \ ,$$

ahol S_i jelöli az *i*-edik tagsági függvényt az *n* dimenziós állapottérben, A_u az *u*-adik tagsági függvényt az egydimenziós akciótérben, $Q_{i,u}$ az egyértékű konklúzió és $\tilde{Q}(s,a)$ pedig a becsült folytonos állapot-akció-érték függvény.

A Q-tanulás közelítő egyenletébe ha behelyettesítjük a nulladrendű Takagi-Sugeno fuzzy következtetést leíró egyenletet az egyértékű következményekre, a következő egyenletet kapjuk eredményül [6], [4]:

$$\begin{aligned} q_{i_1..i_M u}^{k+1} &= q_{i_1...i_M u}^{k} + \prod_{n=1}^{N} \mu_{i_m,n}(s_n) \cdot \mu_u(a) \cdot \Delta \tilde{Q}_{i,u}^{k+1} = \\ &= q_{i_1...i_M u}^{k} + \prod_{n=1}^{N} \mu_{i_m,n}(s_n) \cdot \mu_u(a) \cdot \alpha_{i,u}^{k} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v \in U} \tilde{Q}_{j,v}^{k+1} - \tilde{Q}_{i,u}^{k} \right) \end{aligned} \ , \qquad (1)$$

ahol $q_{i_1...i_M u}^{k+1}$ jelöli az $i_1 \ldots i_M u$ -edik fuzzy szabály egyértékű következményének *k+1*-edik iterációját, ha az S_i állapotban az A_u akciót választjuk, S_j az új megfigyelt állapot, $g_{i,u,j}$ a kapott jutalom az $S_i \rightarrow S_j$ állapot-átmenetre, γ a leértékelési tényező, $\alpha_{i,u}^{k} \in [0,1]$ pedig a lépésköz paraméter. A $\mu_{i_m,n}(s_n) \cdot \mu_u(a)$ a nulladrendű Takagi-Sugeno fuzzy következtetés konklúziójának $q_{u,i}$ szerinti parciális deriváltja.

A fuzzy szabály interpoláció (FRI) alapú Q-tanulás (FRIQ-tanulás) a fuzzy Q-tanulás kiegészítése, a ritka (nem teljes) szabálybázisok alkalmazhatóságának érdekében. Számos FRI módszer lelhető fel a szakirodalomban, a manapság használatos módszerekről egy részletes áttekintést nyújt [2], illetve a gyakorlatban használt FRI alapú alkalmazásokat mutat be [7], [8], [9].

A FIVE (Fuzzy rule Interpolation based on Vague Environment – bizonytalan környezet alapú fuzzy szabály interpoláció) egy alkalmazás orientált FRI módszer (lásd [10], [11]), amely viszonylag alacsony számításigényű és közvetlenül használható egyértékű következményt ad (így konkrét gyakorlati alkalmazás esetén nincs szükség további defuzzifikációs lépésre).

A FIVE FRI és az FQ-tanulás kombinációjából születő módszer előnye, hogy az FQ-tanulás szükségszerűen teljes szabálybázisából kihagyhatóak a kiadódó szabályok. Az FRIQ-tanulást az FQ-tanulás nulladrendű Takagi-Sugeno fuzzy modelljének FIVE FRI-vel való helyettesítésével kapjuk [13]. A nulladrendű Takagi-Sugeno fuzzy következtető modell parciális deriváltját a FIVE FRI modell parciális deriváltjára cserélve kapjuk eredményül a Q-tanulás akció-érték függvény iterációját [13]:

ha $\mathbf{x} = \mathbf{a}_k$ valamely k - ra :

$$q_{i_1 \ldots i_M u}^{k+1} = q_{i_1 \ldots i_M u}^{k} + \Delta \tilde{Q}_{i,u}^{k+1} =$$
$$= q_{i_1 \ldots i_M u}^{k} + \alpha_{i,u}^{k} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v \in U} \tilde{Q}_{j,v}^{k+1} - \tilde{Q}_{i,u}^{k} \right)$$

egyébként : , (2)

$$q_{i_1 \ldots i_M u}^{k+1} = q_{i_1 \ldots i_M u}^{k} + \prod_{n=1}^{N} \left(1 / \delta_{s,n}^{\lambda} \right) / \left(\sum_{k=1}^{r} 1 / \delta_{s,k}^{\lambda} \right) \cdot \Delta \tilde{Q}_{i,u}^{k+1} =$$
$$= q_{i_1 \ldots i_M u}^{k} + \prod_{n=1}^{N} \left(1 / \delta_{s,n}^{\lambda} \right) / \left(\sum_{k=1}^{r} 1 / \delta_{s,k}^{\lambda} \right) \cdot \alpha_{i,u}^{k} \cdot \left(g_{i,u,j} + \gamma \cdot \max_{v \in U} \tilde{Q}_{j,v}^{k+1} - \tilde{Q}_{i,u}^{k} \right)$$

ahol $q_{i_1 \ldots i_M u}^{k+1}$ az $i_1 \ldots i_M u$-edik fuzzy szabály konklúziójának *k+1*-edik iterációja, az S_i állapotból indulva az A_u akciót követően, S_i az új megfigyelt állapot, $g_{i,u,j}$ az $S_i \rightarrow S_j$ állapot-átmenetre kapott jutalom, γ a leértékelési tényező, $\alpha_{i,u}^{k} \in [0,1]$ pedig a lépésköz paraméter.

Az FRIQ-tanulással így lehetővé válik folytonos terek használata a Q-tanulás eredetileg diszkrét állapot-akció tere helyett. A ritka fuzzy szabálybázisok bevezetésével pedig a szabálybázis modell mérete jelentősen csökkenthető a kevésbé fontos szabályok elhagyásával.

2.1. Inkrementális szabálybázis konstrukció

Az interpoláció bevezetésének köszönhetően az FRIQ-tanulás ritka szabálybázis modelljével lehetőség nyílik az állapot-akció-érték függvény inkrementális felépítésére is [14]. A hagyományos fuzzy irányítási rendszerekben megszokott teljes szabálybázisokkal (amelyekben az összes lehetséges szabály szerepel) ellentétben, a módszer kezdetben csak egy minimális méretű (2^{N+1} fuzzy szabály) szabálybázist hoz létre, amiben a fuzzy szabályok az *N+1* dimenziós antecedens (állapot-akció tér) hiperkocka sarkaiban helyezkednek el. A továbbiakban a szabálybázis építési stratégia folyamatosan növeli a kezdeti szabálybázis méretét olyan módon, hogy amennyiben szükség van rá, úgy a megfelelő helyre egy új szabályt helyez be. Abban az esetben, mikor az akció-érték frissítés értéke magas (pl. magasabb mint egy előre definiált érték: ε_Q : $\Delta Q > \varepsilon_Q$) és a legközelebb eső már létező fuzzy szabály is távol van (előre meghatározott értéknél nagyobb a távolság), akkor egy új szabályt illeszt be a legközelebb eső lehetséges helyre. A lehetséges szabály pozíciók egy előre meghatározott stratégia szerint kaphatóak meg, pl. $s_{k+1}=s_k$, $\forall k > i$, $s_{i+1}=\frac{s_i+s_{i+2}}{2}$. Ezzel ellentétben, ha az érték frissítés viszonylag alacsony ($\Delta \tilde{Q} \leq \varepsilon_Q$), vagy a szóban forgó állapot-akció egy már létező szabály közelében van, akkor a szabálybázis érintetlen marad.

Függetlenül attól, hogy került-e be új szabály vagy sem, a konklúziók (Q értékek) a korábban bemutatott FRIQ-tanulási algoritmusnak megfelelően mindig frissülnek. A kapott

akció-érték függvényt így egy olyan ritka szabálybázis fogja modellezni, amiben csak azok a szabályok szerepelnek, amelyek a leginkább szükségesek.

3. Dekrementális szabálybázis redukció

Az inkrementálisan létrehozott szabálybázisban nagy valószínűséggel vannak olyan szabályok is, amelyeknek csak az építési folyamatban volt szerepük, vagy az interpoláció használatának köszönhetően más szabályokból kiadódó szabályok. Ezen szabályok megkeresésére és eliminálására, különböző dekrementális szabálybázis redukciós stratégiák kerültek kifejlesztésre, ezek bemutatása következik az alábbiakban.

Az FRIQ-tanulásban alkalmazott mohó akció választásból eredően [13] a nagyobb Q értékek valószínűsíthetően nagyobb befolyással vannak a rendszer egészének működésére. Így a magas konzekvens értékekkel (itt a Q) rendelkező szabályoknak feltehetőleg nagyobb hatásuk van a rendszerre, illetve a jutalmakra.

Ebből adódóan érdemes lehet olyan redukálási stratégiát választani, ami azokat a szabályokat hagyja el az előzőleg inkrementálisan létrehozott szabálybázisból, amelyeknek alacsony a konzekvens Q értékük (továbbiakban I. stratégia). Ezt a stratégiát követve a redukciós folyamatban a szabálybázisból egyesével kerülnek ki a szabályok (mindig a legkisebb abszolút Q értékkel rendelkező). Minden szabály kivétele után az egész szimulációs folyamat megismétlődik, és ha az így kapott eredmény nem tér el (jelentősen) az eredeti szabálybázissal kapott eredménytől, akkor a vizsgált szabály végleg eltávolításra kerül a szabálybázisból, ellenkező esetben a szabály kardinális szabálynak minősül, így visszakerül a szabálybázisba.

A konkrét problémától függően, az összesített jutalomban megengedhető bizonyos eltérés az eredeti és a módosított szabálybázis között, ha az adott probléma még sikeresen megoldható a redukált szabálybázissal. Hogy mekkora lehet ez az eltérés az a feladat megoldásának a követelményeitől és az adott probléma jellegétől függ. Kis eltérés megengedése a jutalomban jó eséllyel olyan redukált szabálybázist eredményez, aminek használatával hasonló lesz a probléma lépésről-lépésre való megoldása, mint az eredeti szabálybázis használatával. Ellenben ha viszonylag nagy (a konkrét problémához definiált jutalomfüggvénytől függően) a tűréshatár az összesített jutalmak között, akkor a redukált szabálybázis eltérő lépéseket (de helyes eredményt) adhat.

A következő redukciós stratégia (továbbiakban II. stratégia) az előző stratégiához nagyon hasonló, annyiban tér el, hogy a legnagyobb konzekvens értekkel (Q érték) rendelkező szabályokat vizsgálja meg először. Adódhatnak olyan esetek, ahol a végleges szabálybázis különbözni fog az I. és a II. stratégia között, illetve a két stratégia futásidejében lehetnek eltérések.

Egy másik kifejlesztett stratégia (továbbiakban III. stratégia) szabály csoportokat jelöl ki vizsgálatra és eltávolításra, így tömeges szabály eltávolításra ad lehetőséget, ezáltal bizonyos esetekben (a konkrét Q értékektől függően) a redukciós folyamat jelentősen gyorsabb lefutása érhető el, ellenben lehetnek olyan esetek is, ahol ez a stratégia az előzőekhez képest hosszabb futási időt eredményezhet. Ez a módszer először meghatározza a Q értékek teljes tartományát, és a tartomány hosszának a felénél lévő Q értéket tűréshatárként véve osztja fel a szabálybázist két részre. Ezután a nagyobb Q értékű szabályokat tartalmazó szabálybázis kerül kiértékelésre. Ha ez a csökkentett szabálybázis is elegendőnek bizonyul a probléma sikeres megoldásához, akkor ebből a szabálybázisból kiindulva megismétlődik az előző eljárás. Abban az esetben, ha az ideiglenesen csökkentett szabálybázis nem

elegendő (túl sok szabály lett kivéve, tehát a tűréshatár túl tág) a probléma megoldásához, a kivett szabályok visszakerülnek a szabálybázisba. Viszont az előzőleg használt tűréshatár megfeleződik, és ezzel a tűréshatárral ismétlődik meg az ideiglenesen törlendő szabályok meghatározása. Ez a folyamat addig ismétlődik, amíg a tűréshatár elér egy olyan értéket, amikor már az adott szabálycsoport eltávolítható. Abban az esetben, ha már csak 1 szabály marad az adott tűréshatár szerint, és a probléma megoldása így is sikertelen, akkor a szabály „állandó" jelölést kap (így a későbbiekben már nem lesz többször kivételre választva), és a folyamat újraindul egy új tűréshatárral (az előzőleg „állandó"-ra megjelölt szabály ebbe már nem fog beleszámítani).

Az egész folyamat addig ismétlődik, amíg a végső szabálybázisban csak a kardinális szabályok maradnak bent (tehát az „állandó"-nak jelölt szabályok).

Érdemes megemlíteni, hogy a különböző stratégiák különböző végleges szabálybázisokat hozhatnak létre, más-más szabályokkal, de mégis helyes (akár lépésről-lépésre megegyező) megoldást adnak a problémára.

A redukciós folyamat végeztével, a végleges szabálybázisban már csak a legjelentősebb szerepű szabályok fognak szerepelni, másként fogalmazva ez a módszer a rendszer működtető tudását nyeri ki fuzzy szabályok formájában.

3.1. Mintaalkalmazás dekrementális szabálybázis redukcióhoz

Ez az alfejezet egy mintapéldán keresztül mutatja be a fentebb leírt eljárás alkalmazhatóságát, illetve alkalmazásának eredményeit.

A „Cart-Pole", vagy más néven „Reversed Pendulum", azaz fordított inga probléma jól ismert a megerősítéses tanulás témakörében. Egy már létező „Cart-Pole" szimuláció nyílt forráskódú implementációjába került bele a FRIQ-tanulás algoritmusa. Ezt az eredetileg diszkrét térben működő implementációt José Antonio Martin H. készítette MATLAB környezetben [16]. Ebből az implementációból átemelhető a jutalomfüggvény, az akciók hatását leíró függvény, és a megjelenítést megvalósító eljárás. Ezeket a FRIQ-tanulási keretrendszerbe [17] építve és megfelelően paraméterezve elkészíthető a probléma szimulációja.

A szimuláció célja, hogy egy kiskocsi mozgását úgy szabályozza, hogy a rajta csuklóval rögzített, függőleges rudat egyensúlyban tartsa. A szimuláció epizódokon keresztül fut. Egy epizód akkor ér véget, ha a rúd eldől, illetve a kiskocsi falnak ütközik, tehát ha sikertelen a szabályozás, vagy akkor, ha sikerül a rudat egyensúlyban tartani 1000 lépésen keresztül, tehát ha a szabályozás sikeres. A legnagyobb jutalom úgy érhető el, hogy a rúd teljesen függőleges helyzetben van, és a kiskocsi egyenlő távolságra van a falaktól. Sikertelenség esetén pedig a jutalom negatív.

Az állapot-akció-érték függvényt reprezentáló fuzzy szabályok a következőképpen néznek ki:

HA s_1 = $A_{1,i}$ **ÉS** s_2 = $A_{2,i}$ **ÉS** s_3 = $A_{3,i}$ **ÉS** s_4 = $A_{4,i}$ **ÉS** a = $A_{5,i}$ **AKKOR** q = B_i

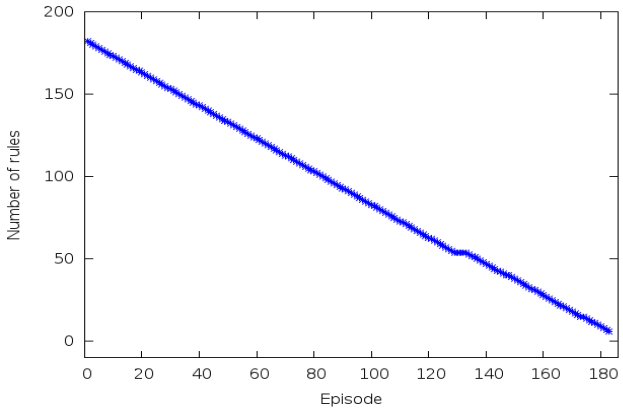
Látható, hogy az állapotteret 4, az akciót pedig 1 változó írja le, ezek a következőek:

s_1 – a kiskocsi helyzete a középponthoz viszonyítva,
s_2 – a kiskocsi sebessége,

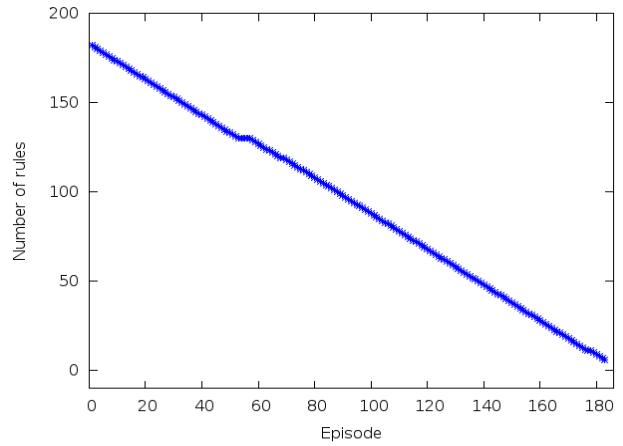
s_3 – a rúd függőlegestől való eltérése,
s_4 – a rúd szögsebessége,
a – a kiskocsi beavatkozó akciója.

Az antecedens részben használt kifejezések az állapotok leírásához a következőek: Negatív (N), Zérus (Z), Pozitív (P), 3° többszörösei [-12°, 12°] tartományban: (N12, N9, N6, N3, Z, P3, P6, P9, P12), és az akciók leírásához: negatívtól pozitívig 0.1 felbontással: AN10-AP10, Z.

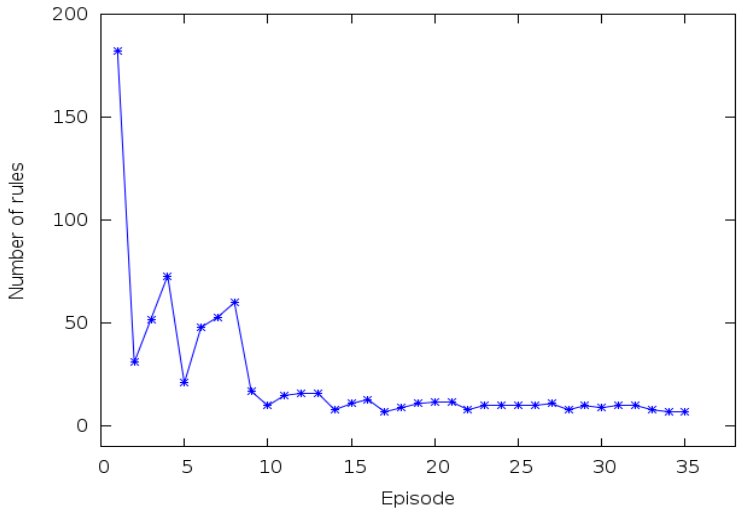
Ennél a példa alkalmazásnál az inkrementálisan építkező módszer egy 182 szabályból álló szabálybázist hoz létre. Ebből a szabálybázisból kiindulva a I. és II. stratégiákat követve ez 183 epizód iterációt jelent (182 minden egyes szabályra, illetve 1 a helyes összesített jutalom és lépésszám meghatározása, bármely szabály eltávolítása nélkül), és nagy eséllyel iterációnként eggyel kevesebb szabály lesz a szabálybázisban, így folyamatosan csökken a számítási igény is. A különböző stratégiák eredménye látható az 1. ábraán, 2. ábra és 3. ábraán.



1. ábra. *A szabályok számának változása az I. stratégiát követve*



2. ábra. *A szabályok számának változása a II. stratégiát követve*



3. ábra. *A szabályok számának változása a III. stratégiát követve, abban az esetben, ha a jutalmak pontos egyezése nem követelmény*

A redukció sikeressége két különböző feltétel szerint került meghatározásra. Először az összesített jutalmaknak szigorúan egyezniük kellett a redukált és az eredeti, inkrementálisan felépített szabálybázis használatával. A másik esetben pedig az összesített jutalmak között bármekkora eltérés lehetett, azzal a feltétellel, hogy összességében a feladatot teljesíti a redukált szabálybázis. Természetesen így a megoldáshoz vezető lépések eltérhetnek a két esetben, de mégis mindkét esetben sikeresek lesznek az epizódok.

A minta alkalmazást a fentiek szerint futtatva a legkisebb szabálybázis az I. és II. stratégia szerint mindössze 7 szabályból áll, továbbá ha megengedett az eltérés a jutalmakban, akkor egy 5 szabályból álló szabálybázis is elegendő a probléma megoldásához. Ezek a szabálybázisok láthatóak a következőekben (1. táblázat és 2. táblázat).

A végleges szabálybázisban található szabályok listája, abban az esetben, amikor pontos jutalom egyezés volt kikötve:

1. táblázat. *A végleges szabálybázis pontos jutalomegyezés esetén*

R#	s_1	s_2	s_3	s_4	a	Q
1	P	Z	Z	P	AP10	1325.1
2	P	Z	N3	N	AN10	1316.5
3	P	Z	Z	N	AN8	1322
4	P	Z	N3	P	AP8	1317.1
5	N	Z	N12	N	AP10	-5251.7
6	P	P	Z	N	AN8	-3100.5
7	P	Z	P12	P	AP4	-6617.7

A végleges szabálybázisban található szabályok listája, abban az esetben, amikor a pontos jutalom egyezés nem volt feltétel:

2. táblázat. *A végleges szabálybázis tetszőleges jutalomeltéréssel*

R#	s_1	s_2	s_3	s_4	a	Q
1	P	Z	Z	P	AP10	1325.1
2	P	Z	N3	N	AN10	1316.5
3	P	Z	Z	N	AN8	1322
4	P	P	Z	N	AN8	-3100.5
5	P	Z	P12	P	AP6	-6446.9

Továbbá a III. stratégia szempontjából jelentős, hogy az egyes stratégiák egymáshoz képest milyen gyorsan fejezik be a redukciós folyamatot. Az alábbi táblázatból tisztán látható, hogy a csoportos szabály elimináció nagyságrendekkel gyorsabb futást eredményez. Ebben az esetben egy olyan szabálybázis az eredmény, amely szintén 7 szabályból áll és csak egyetlen szabályban tér el a I. és II. stratégia által előállított szabálybázistól. Érdekes eredmény továbbá a III. stratégiánál a zéró és a végtelen toleranciával redukált szabálybázisok különbsége: a szabályok száma mindkettőben 7, viszont kizárólag egyetlen közös szabályuk van (lásd a következőekben).

A végső szabálybázist mutatja a III. stratégiát követve zéró jutalomeltérés toleranciával a 3. táblázat.

3. táblázat*. A végső szabálybázis a III. stratégiát követve zéró jutalomeltérés toleranciával*

R#	s_1	s_2	s_3	s_4	a	Q
1	P	Z	Z	P	AP10	1325.1
2	P	Z	N3	N	AN10	1316.5
3	P	Z	Z	N	AN8	1322
4	P	Z	N3	P	AP8	1317.1
5	N	Z	N12	N	AP10	-5251.7
6	N	P	N12	N	AN8	-5038.7
7	**P**	Z	**P12**	**P**	**AP4**	**-6617.7**

A végső szabálybázist mutatja a 4. táblázat a III. stratégiát követve bármekkora jutalomeltérést megengedve (az epizód sikere továbbra is követelmény):

4. táblázat. *A végső szabálybázis a III. stratégiát követve tetszőleges jutalomeltérés esetén*

R#	s_1	s_2	s_3	s_4	a	Q
1	P	N	P12	P	AP10	-6118.1
2	P	N	P12	P	Z	-4012.2
3	P	P	N12	N	AP7	-5140.9
4	P	P	N12	N	AP5	5379
5	P	Z	P12	P	AN1	-5710.3
6	**P**	**Z**	**P12**	**P**	**AP4**	**-6617.7**
7	P	N	P12	P	AN9	-5766.9

A különböző redukciós stratégiák eredményei láthatóak az 5. táblázatban összefoglalva, a Cart-Pole probléma esetén:

5. táblázat. *A különböző redukciós stratégiák eredményei*

Stratégia	*Epizódok száma*	*Szabályok száma*	*Futási idő*
I. 0 diff	183	7	≈3445 s
I. ∞ diff	183	5	≈3442 s
II. 0 diff	183	7	≈3484 s
II. ∞ diff	183	5	≈3432 s
III. 0 diff	63	7	≈256 s
III. ∞ diff	35	7	≈101 s

Az eredményül kapott szabálybázis az eredeti szabálybázishoz képest jelentősen kisebb méretű (lásd 1. táblázat), ember által is könnyen értelmezhető mennyiségű szabályt tartalmaz. A módszer segítségével kinyert működtető tudást tehát a következőképpen lehet értelmezni természetes nyelven:

1. Ha jobbra van a középponttól és nem mozog a kiskocsi és rajta a rúd függőleges, viszont jobbra dől, akkor teljes gőzzel jobbra
2. Ha jobbra van a középponttól és nem mozog a kiskocsi és rajta a rúd kicsit balra áll, de nem dől épp semerre, akkor teljes gőzzel balra.
3. Ha jobbra van a középponttól és nem mozog a kiskocsi és rajta a rúd függőleges, viszont balra dől, akkor erősen balra.
4. Ha jobbra van a középponttól és nem mozog a kiskocsi és rajta a rúd kicsit balra áll és a rudat a lendület jobbra viszi, akkor erősen jobbra.
5. Ha balra van a középponttól és nem mozog a kiskocsi és rajta a rúd erősen balra áll és balra dől, akkor jobbra menni tilos.
6. Ha jobbra van a középponttól és jobbra halad a kiskocsi és rajta a rúd függőleges, viszont balra dől, akkor erősen balra menni tilos.
7. Ha jobbra van a középponttól és nem mozog a kiskocsi és rajta a rúd erősen jobbra áll és a rudat a lendület jobbra viszi, akkor közepes erővel jobbra menni tilos.

4. Összefoglalás

A fuzzy szabály interpoláció alapú Q-tanulás és annak inkrementális szabálybázis konstrukciós módszerének alkalmazása, és a bemutatott dekrementális szabálybázis redukciós stratégiákkal való bővítése hatékony tudáskinyerési eljárást eredményez. A kifejlesztett szabálybázis redukciós stratégiák különböző, de megfelelő megoldásokhoz vezethetnek, így együttes alkalmazásuk vezethet optimális eredményhez. Megfelelően definiált jutalomfüggvény esetén a javasolt módszer képes automatikus tudásfeltárást végezni, tehát meghatározni a rendszert működtető tudást fuzzy szabályok formájában. Továbbá a redukciós eljárásoknak köszönhetően a végső szabálybázisban csak a kardinális szabályok szerepelnek, így ténylegesen a lényeges tudást reprezentáló fuzzy szabályok nyerhetőek ki, amelyek nem csak gép, hanem ember által is közvetlenül kiolvasható alakban szerepelnek. Az eljárás alkalmazhatóságát a bemutatott mintapélda támasztja alá.

5. Köszönetnyilvánítás

A kutatás a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosító számú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt keretében zajlott. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

6. 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] Baranyi, P., Kóczy, L. T., Gedeon, T. D., *"A Generalized Concept for Fuzzy Rule Interpolation"*, IEEE Trans. on Fuzzy Systems, vol. 12, No. 6, 2004, pp. 820-837.

[3] Bellman, R. E.: *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957

[4] 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.

[5] Bonarini, A.: *Delayed Reinforcement, Fuzzy Q-Learning and Fuzzy Logic Controllers*. In Herrera, F., Verdegay, J. L. (Eds.) Genetic Algorithms and Soft Computing, (Studies in Fuzziness, 8), Physica-Verlag, Berlin, D, 1996, pp. 447-466.

[6] Horiuchi, T., Fujino, A., Katai, O., Sawaragi, T.: *Fuzzy Interpolation-Based Q-learning with Continuous States and Actions*. Proc. of the 5th IEEE International Conf. on Fuzzy Systems, Vol.1., 1996, pp. 594-600.

[7] Johanyák, Z.C.: *Survey on Five Fuzzy Inference-Based Student Evaluation Methods*, in I.J. Rudas et al. (Eds.):Studies in Computational Intelligence, 2010, Vol. 313, Computational Intelligence in Engineering, pp. 219-228.

[8] Johanyák, Z. C., Parthiban, R. , Sekaran, G.: *Fuzzy modeling for an anaerobic tapered fluidized bed reactor*, Scientific Bulletin of "Politehnica" University of Timisoara, Romania, Transactions on Automatic Control and Computer Science, vol. 52(66), no: 2, June 2007, pp.67-72.

[9] Johanyák Z. C., Ádámné, M.A.: *Fuzzy Modeling of the Relation between Components of Thermoplastic Composites and their Mechanical Properties*, Proceedings of the 5th International Symposium on Applied Computational Intelligence and Informatics (SACI 2009), May 28-29, 2009, Timisoara, Romania, pp. 481-486.

[10] Kovács, Sz., "*New Aspects of Interpolative Reasoning", Proc. of the 6th. International Conf. on Information Processing and Management of Uncertainty"* in Knowledge-Based Systems, Spain, 1996, pp. 477-482.

[11] 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.

[12] Sutton, R. S., Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, 1998

[13] Vincze, D., Kovács, Sz.: *Fuzzy Rule Interpolation-based Q-learning*, SACI 2009, 5th International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, May 28-29, 2009, ISBN: 978-1-4244-4478-6, 2009, pp. 55-59.

[14] 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.

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

[16] José Antonio Martin H. fordított inga alkalmazása diszkrét térre: (2014. július) http://www.dia.fi.upm.es/~jamartin/downloads/SARSA%20CartPole.zip

[17] A FRIQ-tanulási keretrendszer és a mintaalkalmazás elérhetősége: (2014. július) http://users.iit.uni-miskolc.hu/~vinczed/friq/