FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA

Kolozsvár, 2000. március 24-25.

MEGERŐSÍTŐ TANULÁSI MÓDSZEREK ALKALMAZÁSA INTERNET ROUTING PROBLÉMÁK MEGOLDÁSÁRA

Stefán Péter

Abstract

This paper outlines the fundamental concepts of Reinforcement Learning. Firstly a categorical sorting of learning methods, including RL, will be given, then a short description follows and finally some fundamental application possibilities, such as Internet Routing will be mentioned.

1. Bevezetés

A különféle tanulási módszerek napjainkra kiemelkedő jelentőségre tettek szert. Ezt a tényt igazolja, hogy megnőtt a kereslet az olyan rendszerek iránt, amelyek önálló módon képesek alkalmazkodni változó környezetükhöz, képesek annak struktúrájának, illetve paramétereinek függvényében módosítani saját belső jellemzőiket. Az előbbi, tágabb módon definiált jelenség az emberi tanulást, illetve az emberi megismerési folyamatot modellező rendszer működése.

2. Tanulási módszerek

A tanulási metódusok, külső környezettel való kapcsolat alapján történő osztályozás szerint megkülönböztetünk felügyelt, felügyelet nélküli, illetve megerősítő tanulási módszereket. A környezet, illetve a rendszer kölcsönhatása alapján a legszorosabb a felügyelt tanulás. Ennek sajátossága az, hogy a tanulás folyamán létezik a rendszernek egy olyan környezete, amely a tanulás minden pillanatában meg képes mondani a rendszer helyes viselkedését. Ezt a külső rendszert tanítónak nevezzük. A rendszer viselkedése illetve annak tanító által elvárt viselkedéséből képezhető egy hibajel, ami visszavezethető a

rendszer paramétereinek megváltoztatására. Itt a tanulás iránya kívülről befelé hat, így szokás ezt a fajta módszert instruktív tanítási módszernek is nevezni. A tanító környezet a tanítás minden lépésében utasítja a rendszert a változásra.

A tanulási módszerek másik véglete az önszervező megoldások, melyek esetén a tanuló rendszer nem rendelkezik olyan környezettel, amely meg tudná neki mondani a helyes megoldást. Általában valamilyen attribútum alapján csoportokba rendezi az input mintákat, és esetleges összerendelést végez a csoportjellemző, illetve a csoportba tartozó minták között. Ezek a módszerek képesek az általánosítás illetve a specializáció, mint alapvető tanulási műveletek elvégzésére, de az esetek többségében lassúak, időkorlátos feladatok elvégzésére kevésbé alkalmasak.

Egy harmadik, az előbbi kettő között elhelyezkedő módszer a megerősítő tanulás, amely megpróbálja összegezni azok előnyeit.

3. Megerősítő tanulás

A megerősítő tanulás legkönnyebben az ún. ágens-problémával definiálható. A tanuló rendszert, a kölcsönös analógiának köszönhetően, ágensnek fogjuk nevezni. Az ágens definíciója nem egyszerű dolog, meg kell elégednünk jellemzőinek körülírásával. A két legfontosabb jellemző: az önálló döntéshozatal képessége, valamint a környezetéhez való belülről jövő alkalmazkodás. Mindkét jellemzőről lesz még szó a későbbiekben.

Az ágens feladata az, hogy felderítse környezetét. Tehát, akárcsak más tanulási koncepciók esetén, itt is van egy külső környezet, ami viszont passzívabb a felügyelt tanulórendszerek környezetétől. Nagyon fontos különbség, hogy nem utasítja a rendszert attribútum-, illetve struktúraváltozásra, mindössze a rendszer környezet felé irányuló akcióira ad válaszünterakciót. Az 1. ábra illusztrálja a rendszer környezetével való viszonyát.

1. ábra: Az ágens-probléma

Az ábrán a rendszer és a környezete élesen elkülönül egymástól. A tanulási probléma definiálásánál meg kell tudnunk mondani, hogy mi az, ami a környezethez tartozik, és mi az, ami a rendszerhez. Eredeti definíció szerint minden olyan jellemző a környezethez tartozik, amely a rendszer által nem módosítható. Így lehet az, hogy egy megerősítő tanulással működő robotnak csak az "agya" a tanuló rendszer, azaz az ágens: minden más szenzortechnikai, illetve mozgató elem a környezet része.

A tanulás mechanizmusa az alábbi: a rendszer rendelkezik belső állapotokkal, melyeket egy S halmaz jelöl. ($S = \{s_1, s_2, \dots, s_n\}$) Az S halmaz függhet az időtől, az egyes lépésektől. A rendszer minden egyes s_i állapotban választhat egyet az $A = \{a_1, a_2, \dots, a_m\}$ halmaz elemei közül. Az A halmaz elemeit akcióknak nevezzük. Minden egyes akció választása reakciót vált ki a környezetből. Az egyik reakció, hogy a tanuló rendszer az állapot választása után egy újabb s_j állapotba kerülhet. Ez az állapotváltozás az ábrán a környezetnek van tulajdonítva, mert áttételesen mindig az okozza azok megváltozását. (Egy egyszerű eset, amikor maga az akció közvetlenül váltja ki az állapotváltozást.)

A másik visszajelzett jel pedig az R szimbólummal jelölt skaláris megerősítő jel. Ez jelzi vissza a rendszernek, hogy az adott állapotban kiválasztott akció milyen hatást váltott ki a környezetből. A szám azt méri, hogy "mennyire tetszett" a környezetnek a rendszer viselkedése. Egy nagy pozitív szám jelölheti azt, hogy a környezet méltányolta a választott akciót, egy nagy negatív szám azt, hogy nem tetszett neki, egy nullához közeli érték pedig a környezet "közönyét" fejezi ki. Az már a rendszer saját felelőssége, hogy miként dolgozza fel ezt az információt: figyelmen kívül hagyja-e, avagy hatására valamilyen változást eszközöl saját szerkezetében, paramétereiben. A visszakapott nulla-körüli érték rendszerint nem változtatja meg a tanuló-rendszert, a nagy pozitív, illetve negatív számok rendelkeznek effektív információ-tartalommal. (R értékkészlete megállapodás kérdése. Esetünkben feltételezzük a teljes valós számintervallumot.)

4. A megerősítő tanulás kellékei

Az előbbieken bevezettük a rendszer-állapot, az akció, illetve a megerősítő jel fogalmát. Ahhoz, hogy a megerősítő tanulás megfelelően működjön a rendszernek szüksége van ún. stratégia-függvényre, állapot-érték, illetve akció érték függvényekre. A stratégiai függvény azt mutatja meg, hogy az adott állapotban melyik a legpreferáltabb akció. (Vannak olyan megoldások is, amikor a stratégia az egyes állapotok választási valószínűségét jelenti, ami hasonló funkciókkal bír.) Amikor egy rendszer következetesen minden állapotban az általa legjobbnak vélt akciót választja, azt mohó algoritmusnak nevezzük. A megerősítő tanulás egyik erénye éppen abban áll, hogy a rendszer a visszajelzett R jeltől függően változtatja meg a stratégiáját, minek következtében változhat a "legjobb akció" is. A másik fontos jellemző viszont azt teszi lehetővé, hogy egy bizonyos valószínűségi értéktől függően ne állandóan a mohó algoritmusnak engedjünk, hanem válasszunk olyan megoldásokat is, amelyek, bár rövidtávon nem

preferáltak, hosszabb távon azonban erősebb pozitív reakciót válthatnak ki a környezetből. Ezt a szakirodalom felderítés-kiaknázás közötti egyensúlyként tartja számon, mert minden egyes tanulási lépésben választhatunk két opció között: a már meglévő tudásunkat kamatoztatjuk, vagy újabb lehetőségek után nézünk. Értelemszerű, hogy az előbb említett arányszám sem fix érték, függ a rendszer tapasztalatától is. A tanulás kezdeti szakaszán a felderítési lépések nagyobb aránya, míg a rendszer használatakor a kiaknázó lépések a jellemzők. A tudás kiaknázásának, illetve az újabb tudás megszerzésének együttes, egy működési cikluson belüli használata lehetővé teszi, az ágens sokoldalú felhasználását.

Az állapot-érték, illetve az akció-érték (pontosabban állapot-akció-érték) függvények hosszú távú becslések. A tanulórendszer feladata, hogy a környezettől visszakapott megerősítő jelet hosszú távon maximalizálja. Azonban az akció választásának pillanatában ez a jel még nem áll a rendelkezésére, mindössze annak egy várható értéke. Ezek a várható értékek az érték-függvények. Az állapot-érték függvény azt mutatja meg, hogy mennyi a várható honorárium az egyes állapotokba jutásnak, míg az akció-érték függvény arra ad becslést, hogy a kiválasztott akciót hogyan fogja a környezet értékelni. Az érték függvények nemcsak rövidtávú (egy lépéses) becslést tesznek lehetővé, hanem azt is, hogy a több lépés múlva következő nagy **R** értékű állapotot "előrevetítse" a jelen állapotra.

5. Lehetséges alkalmazások

Az előzőekben megalkotott modell keretei közé számos gépészeti, illetve informatikai probléma helyezhető. Az Internet routing probléma, adaptív robot vezérlés, változó környezetben történő optimum-keresés, illetve logisztikai disztribúciós feladatok megoldása mind ebbe a tárgykörbe sorolható.

6. Felhasznált irodalom

- [1] P.STEFÁN, L.MONOSTORI, F.ERDÉLYI: Using symbolic and sub-symbolic methods for solving problems difficult to analyze, MicroCAD'99 University of Miskolc
- [2] R.SUTTON, A.BARTO: Reinforcement learning

Név: Stefán Péter

Munkahely: Miskolci Egyetem, Alkalmazott Informatika Tanszék

Telefon: +36 46 565 111 mellék: 19-52