# YSA'lar, Pekiştirimli Öğrenme ve Beklenti

Uğur Halıcı
Bilgisayarla Görme ve Yapay Sinir Ağları Araştırma Laboratuvarı
ODTÜ Elektrik ve Elektronik Mühendisliği Bölümü

İletişim:
Bilgisayarla Görme ve Yapay Sinir Ağları Araştırma Laboratuvarı
Elektrik ve Elektronik Mühendisliği Bölümü
06531, ODTÜ, ANKARA
Telefon: (+90) 312 210 2333
Fax: (+90) 312 210 1261
e-posta: halici@metu.edu.tr
http://vision1. eee. metu. du. tr/~halici/

## YSA'lar, Pekiştirimli Öğrenme ve Beklenti

**ÖZET** *Yapay Sinir Ağları (YSA'lar), kendisine sunulan örnekler üzerinden öğrenebilen akıllı sistemlerdir. Günümüzde bir çok uygulamada başarı ile kullanılan McCullough-Pitts modeline dayalı yapay nöronlar gerçek nöronların davranışını modellemekten oldukça uzaktır. Sinyallerin sabit sinyal seviyeleri yerine uyartılar halinde iletildiği Rassal Sinir Ağı (RSA) modelindeki yapay nöronlar, biyolojik nörona daha benzer biçimde davranır. Pekiştirimli öğrenme, yapay sistemlerin eğitiminde kullanılan öğrenme yöntemlerinden biridir ve hayvanlardaki araçsal koşullama ile yakından ilgilidir. Bu makalede RSA'ların pekiştirimli öğrenmesi konusu ele alınmış ve ödül beklentisini göz önüne alan öğreme kuralı sunulmuştur.*

***Anahtar Kelimeler:*** *Yapay sinir ağları, Rassal sinir ağı modeli, Pekiştirimli öğrenme*

## ANNs, Reinforcement Learning and Expectation

**ABSTRACT** *Artificial Neural Networks (ANN) are the intelligent systems that can learn through the samples presented to them. The artificial neurons based on McCullough-Pitts model are used successfully in several applications today, but they are far away from modelling the behaviour of real neurons. The Random Neural Network (RNN) model, in which signals travel as voltage spikes rather than as fixed signal levels, represents more closely the manner in which signals are transmitted in biological neural networks. Reinforcement learning is one of the approach used in training artificial systems and closely related to instrumental conditioning in animal learning. In this paper, reinforcement learning of RNN is considered and a training rule that considers the expectation of reward is presented.*

***Keywords:*** *Artificial neural networks, Random neural network model, Reinforcement learning*

### GİRİŞ

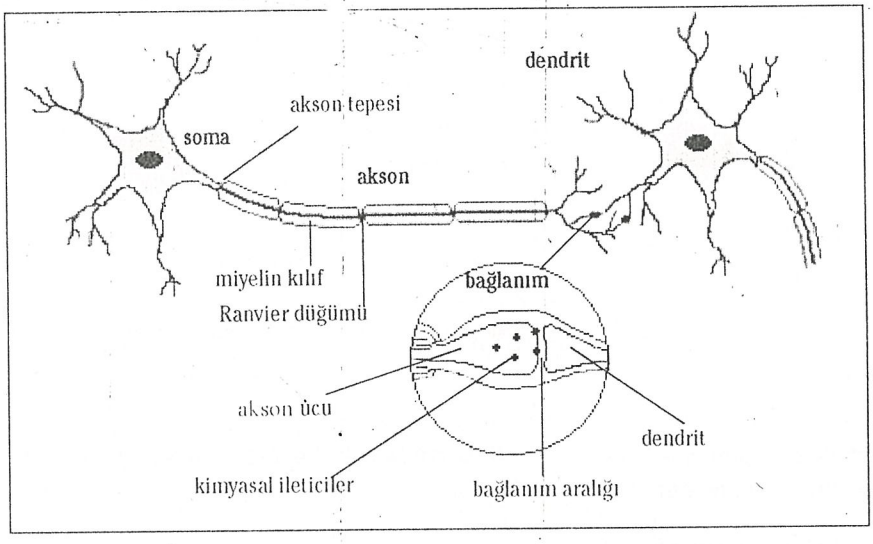
Biyolojik nöronlardan esinlenilerek geliştirilen YSA'lar bu bir çok uygulamada başarı ile kullanılmaktadır. Örüntü (optik karakter, ses, parmakizi vb.) tanıma, görüntü iyileştirme, fonksiyon yaklaştırma, durum tahmini, eniyileme, akıllı kontrol gibi konular bu uygulamalardan bazılarıdır. YSA'lar, yapay nöron adı verilen işlem elemanlarının birbirlerine bağlanmasıyla oluşur.

Biyolojik sinir sisteminde, gönderici nörondaki uyartıların (nerve pulse) alıcı nörondaki uyarımı, akson ucunda bulunan özel kimyasal ileticilerin sinir bağlanım (synapse) aralığına dökülmesiyle oluşan oldukça karmaşık elektro-kimyasal bir olaydır (Şekil 1). Bunun alıcı nörondaki etkisi, nöron gövdesindeki (soma) dereceli potansiyelin (graded potential) artması ya da azalması biçiminde ortaya çıkar. Eğer dereceli potansiyel bir eşik değerine ulaşabilirse nöron ateşlemeye başlar. 1943'te McCulloch ve Pitts tarafından önerilen yapay nöron modelinde yaratılmaya çalışılan işte bu özelliktir. Şekil 2'te gösterilen bu nöron modeli, üzerindeki bazı ufak değişikliklerle bu gün hala yapay sinir ağlarında yaygın bir biçimde kullanılmaktadır. Şekilde verilen yapay nöronda $u_1$, $u_2$, $u_N$ ile gösterilen N tane giriş bulunmaktadır. Bu girişleri nörona bağlayan her bir hatta $w_1$, $w_2$, , $w_N$ ile gösterilen bağlanım kuvveti atanmıştır. Yapay nöron modelindeki bağlanım kuvvetleri, biyolojik nöronlardaki bağlanımların etkileme gücüne karşılık gelir. Negatif bağlantı değerleri ketleyici bağlanımları temsil ederken, pozitif değerler uyarıcı bağlanımları temsil ederler. Bir nöronun ateşleyip ateşlememesini belirleyen eşik değeri yapay nöronda genellikle * ile gösterilir ve nöron toplam uyarımı aşağıdaki denklemle ifade edilir:

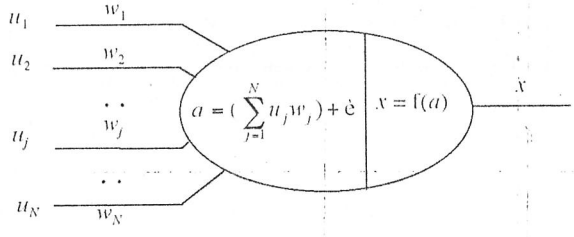
$$a = \left(\sum_{j=1}^{N} w_j u_j\right) + \theta \qquad (1)$$

Biyolojik nörondaki ateşleme sıklığı ile ilişkilendirilerek, nöronun x ile gösterilen çıkış değeri nöron toplam uyarımının bir fonksiyonu olarak yazılır:

$$x = f(a) \qquad (2)$$



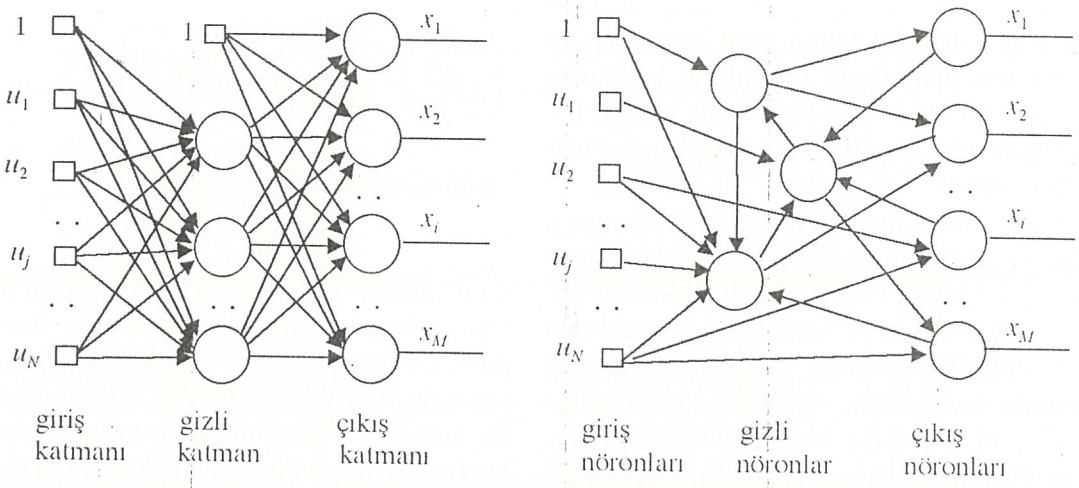

**Şekil 1.** *Tipik bir nöron*




**Şekil 2.** *Yapay nöron*

Özgün McCulloch-Pitts modelinde, çıkış fonksiyonu f(a) bir eşik fonsiyonu olarak önerilmiştir. Ancak doğrusal, yokuş, sigmoid fonksiyonları da yapay sinir ağlarında yaygın olarak kullanılmaktadır. Çeşitli uygulamalarda kullanılan YSA'lar, bir çok nöronun birbirlerine çeşitli yapılarda bağlanmasıyla oluşturulmaktadır. Aralarındaki bağlantıların yapısına göre bu ağlar Şekil 3'te gösterildiği gibi çeşitli sınıflara ayrılırlar (YSA hakkında daha ayrıntılı bilgi için bkz Halıcı 2000-a). Mc-Culloch-Pitts modeline dayalı yapay nöronlar kullanılarak, öğrenebilme yeteneğine sahip akıllı sistemler başarıyla geliştirilmiştir. Ancak, sinyal iletiminin nöron çıkışlarında skalar değer alan bir çıkış fonksiyonu temsil edildiği böyle bir yaklaşım biyolojik nöron davranışını modellemek için fazlaca basittir. Biyolojik nöronlarda sinyaller, sabit sinyal seviyeleri yerine akson üzerinde birbiri ardınca dizilen uyartılar aracılığıyla iletilmektedir.

Hodgkin-Huxley modelinde, membran sığası, membran direnci, iyon kanalları gibi parametreler gözönüne alınarak membran potansiyeli hesaplanmakta ve akson üzerinde ilerleyen uyartılar biyolojik nörondakine çok daha benzer bir biçimde modellemektedir. Ancak bu model, YSA'larda kullanılmak için fazlaca karmaşıktır. Hodgkin-Huxley modelindeki kadar karmaşık olmasa da, nöron sinyallerindeki bilginin darbeler (pulse) ve/veya darbeler arası zamanla kodlandığı bazı uyartılı yapay nöron modeleri son yıllarda geliştirilmiş ve bu modeller YSA ile ilgili araştırmalarda hakettiği yeri almıştır. RSA, nöron giriş ve çıkışındaki sinyallerin darbelerle kodlandığı uyartılı YSA modellerinden biridir (Gelenbe 1989). Beyinde bilginin ne şekilde saklandığı ve bilgiye nasıl erişildiği henüz tam olarak bilinmemektedir. Ancak bu konuda yapılan deneysel çalışmalarda, belirli uyarıların düzenli olarak uygulandığı nöronların yapısında bazı değişikliklerin meydana geldiği gözlenmiştir. Düzenli uyarılar karşısında oluşan en önemli değişiklikler, sinir bağlanımlarının elektriksel ve kimyasal özelliklerinde ortaya çıkmaktadır. Örneğin, sinir bağlanım aralığına dökülen kimyasal iletici miktarı öğrenme ile




**Şekil 3.** a) *Çok katmanlı ve ileri beslemeli ağ, b) Katmansız ve geri beslemeli ağ*

artmakta ya da eksilmekte veya bağlanım sonrası nöronun iletici nöronlara tepkisi değişebilmektedir. Sonuçta öğrenme, bağlanım sonrası nörona ulaşan uyartıların, bu nöronun dereceli potansiyelinin eşik değerine ulaşmasındaki önemini değiştirmekle ilişkilidir. YSA'larda bu durum, bağlanım kuvvetlerini ve eşik değerini değiştirmek suretiyle modellenmekte ve böylece YSA'lar öğrenme ve hatırlama yeteneği kazanmaktadır.

Öğrenme sırasında, bir öğrenme kümesinden alınan örnekler YSA'ya uygulanarak ağdaki bağlanım kuvvetleri yinelemeli (iterative) bir şekilde değiştirilir. Bağlanım kuvvetlerinin ne kadar değişeceği seçilen bir bağlanım kuvveti değiştirme (BKD) kuralına göre belirlenir ve BKD kuralı mutlaka bir öğrenim algoritması çerçevesinde uygulanır. Bu öğrenim algoritmaları en genel anlamda üç kategoriye ayrılır: Denetmenli öğrenme (supervised training), pekiştirimli öğrenme (reinforcement training) ve kendiliğinden organizasyon (self-organization).

### Pekiştirimli Öğrenme

Pekiştirimli öğrenme, eğitmenli öğrenmeye benzer biçimde olur, ancak ağ çıkışında olması gereken değerler ağa doğrudan bildirilmez, bunun yerine ağın ne kadar iyi bir çıkış ürettiğine ilişkin bir değer bildirilir. Pekiştirimli öğrenmede, YSA'ya ne yapması gerektiği doğrudan söylenmemekte, bunun yerine çeşitli durumlar YSA tarafından denenerek en büyük getirisi olan çıkış değerlerinin ağın kendisi tarafından bulunması gerekmektedir. YSA'lardaki pekiştirimli öğrenme, hayvanlardaki araçsal koşullama (instrumental conditioning) ile yakından ilgilidir. Araçsal koşullamada, organizma öğrenme ortamında aktif bir rol üstlenmektedir. Bu tür bir öğrenme, organizmanın yaptığı davranışlarına karşılık çevreden aldığı yanıtlara göre kendi davranışlarını ayarlamasına izin verir. Eğer bir davranış hoşa gidecek bir sonuç yaratıyorsa, bu davranış daha sık gösterilmeye eğilim taşır. Diğer yandan, eğer bir davranış hoşa gitmeyecek bir sonuç yaratıyorsa daha seyrek gösterilmeye eğilim taşır. Genel olarak, hoşa giden sonuçlar ödül (reward), hoşa gitmeyen sonuçlar ise ceza (punishment) olarak adlandırılırlar (Carlson 1977, Hulse, Egeth, Deese 1980). Deneme-yanılma ile en iyi davranışın aranması, pekiştirimli öğrenmenin önemli özelliklerinden biridir. Pekiştirimli öğrenmede, bilineni işletme (exploidation) ve bilinmeyene açınsama (exploration) olarak adlandırabileceğimiz bir çatışma bulunmaktadır. Bunu daha açacak olursak 1) sistem tarafından daha önce yapılmış

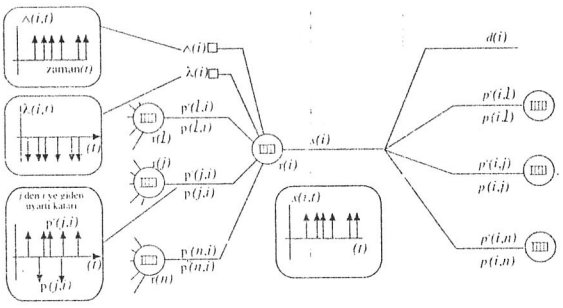
eylemlerin getirileri hakkında halihazırda edinilmiş bilginin kullanımı, yani iyi getirisi olan eylemlerin yapılması yönündeki arzu ve 2) ileride daha iyi bir seçim yapabilmek üzere, iyi bilinmeyen eylemlerin getirileri hakkında daha fazla bilgi elde edinebilme, yani getirisinin ne olduğu pek bilinmeyen eylemlerin yapılması arzusu birbiri ile çatışır. Bu durum, pekiştirimli öğrenmedeki bir zayıflıktır (Narendra, Thathachar, 1989).

Yalnız ödüle dayalı pekiştirimli öğrenme, bilineni işletmeyi desteklerken bilinmeyene açınsamayı önler. Dolayısıyla, bu yönteme göre öğrenen bir sistem, öğrenilmiş eylemlerden birine takılarak kendini değiştirme yeteğini kaybeder ve çevrenin değişmesi ile ortaya çıkacak yeni koşullara kendisini uyarlayamaz. Bu problemin bir çözümü, cezanın BKD kuralına sokularak bilinmeyene açınsamanın desteklenmesidir. Ancak, böyle bir yaklaşım, çevrede hiç bir değişim olmasa bile, en iyi eyleme tümüyle yakınsamayı engeller.

Pekiştirilmeyen ilişkilendirmelerin (association) öneminin azalmasına karşılık gelen sönüm (extinction), hayvanlardaki öğrenmede çok iyi bilinen bir özelliktir. Sönüm sayesinde, sistemin artık geçerli olmayan eski ilişkilendirmeleri unutarak yeni ilişkilendirmeleri öğrenebilmesine olanak doğar. Bu durum, değişen koşullara uyum sağlamaya yaradığı için, yaşayan organizmaların varlıklarını sürdürebilmeleri açısından büyük önem taşımaktadır (Hulse, Egeth, Deese 1980).

### Rassal Sinir Ağı Modeli

Biyofiziksel nöron davranışının basitleştirilmiş bir gösterimi olan RSA modeli, (Gelenbe 1989)'da önerilmiş ve (Gelenbe 1990)'da genişletilmiştir (bkz. Şekil 4). RSA modelinde, N tane nöron pozitif ve negatif uyartıları (pulse) bağlanımlar üzerinden birbirlerine iletmektedir. Bu modeldeki nöronlar diğer nöronlardan uyartı aldıkları gibi ağın dışındaki kaynaklardan da uyartılar alabilirler. Dışsal kaynaklardan üretilen pozitif ve negatif uyartılar (exogenous input), $\wedge(i)$ and $\lambda(i)$ hızına sahip iki Poisson süreci ile modellenmektedir. RSA'daki pozitif uyartılar, uyarıcı bağlanımlardan ulaşan uyartıları temsil ederken, negatif uyartılar ise ketleyici bağlanımlardan ulaşan uyartılara karşılık gelir. Her bir $i$ nöronunun $t$ anındaki dereceli potansiyeli $k_t(i)$ ile gösterilmekte ve bu potansiyel o nörona ulaşan pozitif ve negatif uyartıların sayısına göre belirlenmektedir. Nörona ulaşan her bir pozitif uyartı, $k_t(i)$ değerini bir artırmakta, ulaşan her bir negatif uyartı ise bu değeri bir azaltmaktadır; ancak eğer $k_t(i)=0$ ise bu durumda ulaşan negatif uyartılar bunu daha fazla azaltamaz. Ayrıca, nöron ateşlediği sırada üretilen her bir uyartı da, $k_t(i)$ değerini aynı şekilde bir azaltır. Herhangi bir anda, eğer $k_t(i)$ pozitif ise nöron ateşleyebilir. Ateşleme sırasında, uyartılar sabit hızlı üssel dağılıma (exponential distribution of constant rate) göre rassal olarak üretilmektedir. Ateşleme hızı $r(i)$ ile gösterilmektedir. Üretilen bu uyartılar, ağdaki diğer nöronlara göderilmekte ya da ağın dışına çıkarak yok olmaktadır. Herhangi bir $i$ noronundan çıkan bir uyartı, olasılığı ile $j$ nöronuna pozitif bir uyartı olarak veya $p^+(i,j)$ olasılığıyla negatif bir uyartı olarak iletilir ya da $d(i)$ olasılığı ile ağın dışına çıkarak yok olur. RSA'da, $w^+(i,j)=r(i)p^+(i,j)$ değeri, $i$ ve $j$ nöronları arasındaki uyarıcı bağlanım kuvvetine, $w^-(i,j)=r(i)p^-(i,j)$ ise ketleyici bağlanım kuvvetine karşılık gelir. $N$ ağdaki toplam nöron sayısı, $N_i$ ise i nöronunun bağlı komşu-nöronların sayısı olsun. Her bir i nöronu, $1\leq i \leq N$ için, $p(i,j)=p^+(i,j)+p^-(i,j)$ olasılıklarının j üzerinden toplamı




**Şekil 4.** *Rassal Nöron Modeli*

$$\sum_{j=1}^{Ni} p(ij) + d(i) = 1 \qquad (3)$$

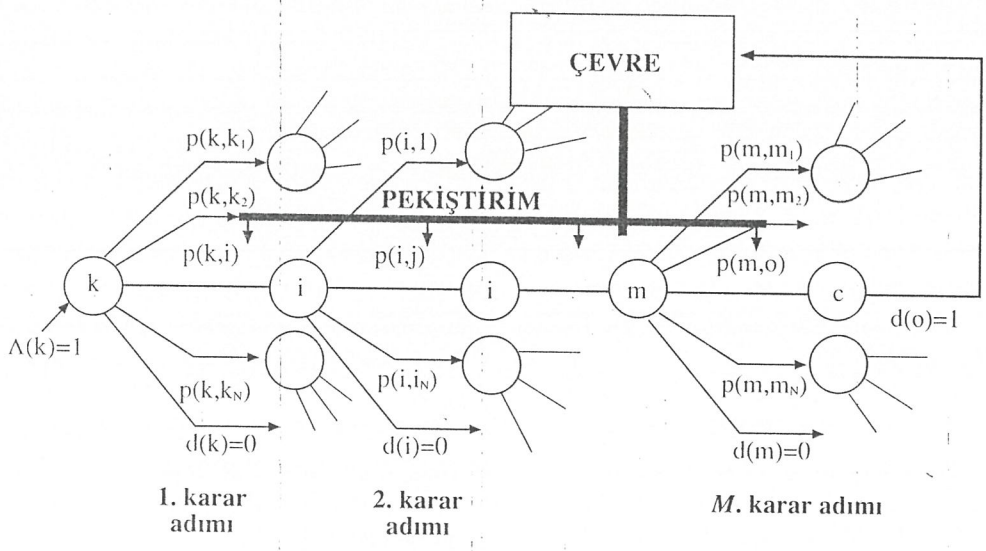
bağıntısını sağlanmalıdır, ayrıca $0 \le p^{+}(i,j) \le 1$ ve $0 \le p^{-}(i,j) \le 1$ olmalıdır.

Yatışkın durumda (steady state), dereceli potansiyellerin pozitif olma olasılıklarının analitik olarak nasıl hesaplanacağı (Gelenbe, 1989)'da gösterilmiştir, ancak YSA'ların yatışkın durum davranışı bu makalenin kapsamı dışındadır.

RSA'ların, (Gelenbe 1989)'da önerilmesinden bu yana, RSA modeline dayalı bir çok uygulama geliştirilmiştir (detaylı liste için Gelenbe 2000'e bakınız). Gradyan azalımına dayalı bir denetmenli öğrenme algoritması (Gelenbe 1993)'de önerilmiştir.

Bulunduğu bir çevrede yaptığı bir eylem (action) dizininden elde ettiği pekiştirime göre öğrenen bir sistem düşünelim. Bu sistemin $n$. bir deneyim sırasında yaptığı $M$ eylemden oluşan eylem dizinini $a_n = \langle a_n(m)\ m=1..M \rangle$ ile gösterelim. Bu an dizinindeki her bir $a(m)$ eylemi, sistemin içinde bulunduğu çevreyi etkilemekte ve sistemin $s_i(m-1)$ durumundan $s_j(m)$ durumuna geçmesine neden olmaktadır. Herhangi bir durumdayken yapılabilecek eylem çeşitleri, o anki durum $s_i(m)$ ile doğrudan ilişkilidir. Öğrenen sistem herhangi bir son duruma $s_f(M)$, eriştiğinde an eylem dizini tamamlanmış olur ve sistem çevreden yaptığı eylemlere karşılık gelen bir $R_n(a_n)$ pekiştirimi alır. Çevreden alınan pekiştirimin miktarı gerçekleştirilen $a_n$ eylem dizininine göre rassal olarak belirleniyor olsun. Bir eylem dizininin tamamlanarak çevreden ilgili pekiştirimin alınması bir deneyime karşılık gelir. Böyle bir sistem için pekiştirimli öğrenimin amacı, alınacak ödülün beklenen değerini (expected value) en çoğa çıkaran veya cezanın beklenen değerini en aza indiren eylem dizinini deneyimler sonucunda bulmaktır.

RSA'ların yukarıda açıklandığı biçimdeki bir pekiştirimli öğrenmede nasıl kullanılabileceği daha önceki bir çalışmamızda sunulmuştu (Halıcı 1997). Ardışıl eylemler dizinini eniyilemek amacıyla kullanılacak RSA yapısı, Şekil 5'de gösterilmiştir. Sistemdeki her bir nöron olası bir duruma karşılık gelmektedir. Bir nörondan dışarı doğru uzanan her bir bağlantı, değişik bir eylemin kararını temsil etmektedir. Ağda bir başlangıç nöronu, belirli sayıda sonuç nöronları bulunmaktadır, diğerleri ise ara




**Şekil 5**. Pekiştirimli öğrenme için önerilen RSA

nöronlardır. Başlangıç nöronu henüz hiçbir eylemin yapılmadığı duruma karşılık gelirken, sonuç nöronları sistemin çevreden pekiştirim aldığı durumlara karşılık gelmektedir. Pekiştirimli öğrenmede kullanılacak RSA parametreleri aşağıdaki gibi seçilir ve bu değerler öğrenme sürecinde sabit kalır :

(4)

Başlangıç nöronu: $r(i)=1,\ \Lambda(i)=\Lambda,\ \lambda(i)=0,\ d(i)=0,\ p(i,j)=0$

Ara nöronu: $r(i)=1,\ \Lambda(i)=0,\ \lambda(i)=0,\ d(i)=0,\ p(i,j)=0$

Sonuç nöronları: $r(i)=1,\ \Lambda(i)=0,\ \lambda(i)=0,\ d(i)=1,\ p(i,j)=0$

Ağdaki tüm $p+(i,j)$ olasılıkları başlangıçta 1/Ni değeri alır ve bu değerler öğrenme sırasında en iyi eylem dizinini belirlemek üzere değişir. Tüm ağda $p'(i,j)=0$ seçildiğinden, $p(i,j)=p+(i,j)$'dir. Bundan sonra basitlik amacıyla $p'(i,j)$ yerine $p(i,j)$ kullanılacaktır. Yalnızca başlangıç nöronu için $\Lambda(i)=\Lambda$, diğerleri için $\Lambda(j)=0$ seçildiğinden, uyartılar sadece başlangıç nöronunda yaratılır, diğer nöronlardan geçerek sonuç nöronlarından birine ulaşır. Herhangi bir *i* nöronuna gelen bir uyartının hangi komşu j nöronuna gideceği, $p(i,j)$ bağlanım olasılık değerlerine göre belirlenmektedir. Dolayısı ile komşu nöronlar arasında en büyük bağlanım olasılık değerine sahip olan nöron en fazla seçilme olasılığına sahip demektir. Sonuç nöronlarından birine ulaşıldığında, burada $d(i)=1$ olduğundan, uyartı buradan çevreye giderek yok olur. Çevre yanıt olarak, yapılan eylemler dizininde yer alan eylemlerin başarısına göre bir pekiştirim üretir. Bu pekiştirim miktarı gözönüne alınarak, yalnızca eylem dizininde yer alan nöronların bağlanım kuvvetleri önceden seçilen bir BKD kuralına göre değiştirilir.

Bir uyartının başlangıç nöronundan başlayarak bir sonuç nöronuna kadar ulaşıp çevreden bir pekiştirimin alınması bir deneyime karşılık gelir. Öğrenme, artarda gelen deneyimlerle devam eder. Her bir deneyimden sonra bağlanım kuvvetleri bir miktar değişir. Belirli bir aşamadan sonra bağlanım kuvvetleri artık belirli bir dengeye oturur ve fazlaca değişmez. Bu durum RSA'nın öğrenmesinin tamamlandığını gösterir.

**Ödülün içsel beklentisini kullanan BKD kuralı**

Öğrenmede uygulanacak BKD kuralının seçimi çok önemlidir. İyi bir BKD kuralı ile deneyimler tekrarlandıkça iyi pekiştirim alan eylemlere karşılık gelen nöronların bağlanımları kuvvetlenirken diğer bağlanım kuvvetlerinin zayıflaması gerekir. Ayrıca sistem çevre değişimlerine uyum gösterebilmelidir. Eğer öğrenilen bir eylem dizini değişen çevre koşullarından dolayı artık başarısız kalıyorsa bu dizin unutulmalı, bunun yerine yeni çevre koşullarında başarılı olacak yeni bir eylem dizini deneyimle öğrenilmelidir. Yalnız ödüle dayalı (Ö-BKD) kuralı, öğrenen otomaton (learnig automata) adı verilen sistemin eğitiminde yaygın olarak kullanılan bir BKD kuralıdır (Narendra Thathachar 1989). Önceki bir çalışmamızda Ö-BKD kuralı RSA için uyarlanarak, bir hedefe ulaşmada toplam gideri (cost) eniyileyecek, örneğin bir labirentte hedefe en kısa yoldan ulaşmayı sağlayacak, ardışıl eylemlerin seçiminde kullanılmıştır (Halıcı 1997). Ö-BDK kuralı ile öğrenen RSA, durağan (stationary) çevrelerde çok iyi başarım göstermektedir. Ancak, durağan olmayan çevrelerde, yalnız ödül ile öğrenmenin bilinmeyene açınsamayı engellemesi nedeniyle, sistem önceden öğrenilen eylem dizinine takılı kalmakta ve değişen durumlara uyum sağlamakta başarısız olmaktadır (Halici 1997). Bu problemin üstesinden gelmek için Ö-BKD kuralı, ödülün içsel beklentisini kullanan bir BKD (ÖB-BKD) kuralı önerilerek, ödül/cezaya dayalı olacak şekilde genişletilmiştir (Halici 2000-b). İçsel beklentiyle pekiştirimli öğrenme olarak adlandırdığımız bu yeni yaklaşımda, gerçekleştirilen bir eylem dizini için elde edilen ödül miktarı beklentinin altında olmadığı sürece ödülle öğrenme gibi davranılmakta, ödül miktarının beklentinin altında olduğu durumlarda ise ceza ile öğrenme uygulanarak diğer olası eylem dizinlerine açınsama sağlanmaktadır.

Aşağıda çok adımlı karaların öğrenilmesinde kullanımak üzere önerdiğimiz ÖB-BKD kuralı verilmiştir.

$$p_{n+1}(i,j)=\begin{cases} p_n(i,j)+\eta^+(R_n^+(\mathbf{a}_n)-\mathbf{R}_{n,\beta}^+)(1-p_n(i,j)) & for\ (i,k_l)\in\mathbf{a}_n,\ j=k_l,\ R_n^+(\mathbf{a}_n)>\mathbf{R}_{n,\beta}^+ \\ p_n(i,j)-\eta^-(\mathbf{R}_{n,\beta}^+-R_n^+(f))p_n(i,j) & for\ (i,k_l)\in\mathbf{a}_n,\ j=k_l,\ R_n^+(\mathbf{a}_n)\le\mathbf{R}_{n,\beta}^+ \\ p_n(i,j)-\eta^+(R_n^+(\mathbf{a}_n)-\mathbf{R}_{n,\beta}^+)p_n(i,j) & for\ (i,k_l)\in\mathbf{a}_n,\ j\neq k_l,\ R_n^+(\mathbf{a}_n)>\mathbf{R}_{n,\beta}^+ \\ p_n(i,j)+\eta^-(\mathbf{R}_{n,\beta}^+-R_n^+(\mathbf{a}_n))(\frac{1}{N_i-1}-p_n(i,j)) & for\ (i,k_l)\in\mathbf{a}_n,\ j\neq k_l,\ R_n^+(\mathbf{a}_n)\le\mathbf{R}_{n,\beta}^+ \\ p_n(i,j) & for\ i\notin\mathbf{a}_n \end{cases} \quad (5)$$

Yukarıdaki ifadede, $n$ deneyim numarasını, $k_l$ ise $i$ nöronundayken seçilen eylemi, yani seçilen komşu nöronu göstermektedir. $\eta^+$ ve $\eta^-$ ödül ve ceza ile öğrenme katsayılarına karşılık gelen küçük pozitif sabitlerdir. $R^+_n(a_n)$, deneyiminde seçilen an eylem dizini için çevreden elde edilen ödülü, $R^+_{n,\beta}$ ise o zamana kadar olan deneyimlerde çevreden alınan ödüllere göre oluşan beklentiyi temsil etmektedir. Yukarıda verilen kurala göre, eğer $R^+_n(a_n)\geq R^+_{n,\beta}$, olursa $R^+_n(a_n)-R^+_{n,\beta}$, etkin ödül miktarını belirlemekte ve deneyim sırasında yapılmış olan $k_l$ eylemleri ile ilgili bağlanım olasılıkları etkin ödül miktarı gözönüne alınarak artırılmakta, diğer bağlanım olasılıkları ise azaltılmaktadır. Dolayısı ile başarılı eylem dizininin görülme sıklığı artarken diğer dizinlerin sıklığı azalacaktır. $R^+_n(a_n)<R^+_{n,\beta}$ olduğu durumda ise $R^-_n(a_n)=R^+_{n,\beta}-R^+_n(a_n)$ etkin ceza miktarını belirlemektedir. Deneyim sırasında gerçekleştirilmiş $k_l$ eylemleri için olan bağlanım olasılıkları azaltılmakta (diğer bağlanım olasılıkları artırılmakta) ve bu eylem dizinin ileride daha az olasılıkla yapılması sağlanmaktadır. Ödül beklentisi $R^+_{n,\beta}$ başlangıçta, yani n=0 olduğu zamanda, $R^+_{0,\beta}=0$ olarak seçilmekte ve her deneyimde aşağıdaki ifadeye göre yenilenmektedir:

$$\mathbf{R}^+_{n+1,\beta}=(1-\beta)\,\mathbf{R}^+_{n,\beta}+\beta R^+_n(\mathbf{a}_n) \quad (6)$$

burada $\beta$ küçük pozitif bir sabittir ve $\min(\eta^+,\eta^-)\leq\beta\leq\max(\eta^+,\eta^-)$ olacak şekilde seçilir.

## SONUÇLAR

ÖB-BKD kuralı ile tek adımlık eylemleri öğrenenen bir RSA'da eylemlerin bağlanım olasılıklarının öğrenme tamalandığında hangi değerleri alacağı (Halıcı 2000-b)'de analitik olarak gösterilmiştir. Bu analitik sonuçlara göre, her bir eylemle ilgili bağlanım olasılıkları ile bu eylemlerin aldığı ödüllerin beklenen değerleri aynı büyüklük sırasında oluşur, yani en fazla ödül getiren eylem en fazla yapılmaktadır. Ayrıca benzetim programları aracılığıyla ÖB-BKD kuralı ile öğrenen RSA'nın davranışı tek adımlı kararlar için deneysel olarak gözlenmiş ve (Naredra, Thathachar 1989)'da incelenen Dogrusal Ödül-Ceza BKD (DÖC-BKD) kuralının sonuçları ile karşılaştırılmıştır. Deneylerde öğrenme fazından sonra çevre koşullarının değiştirildiği bir sönüm fazı kullanılmış, böylece öğrenen sistemin değişen çevre koşullarına uyumu da ayrıca incelenmiştir. Bu deneylerde, ÖB-BKD kuralının sonuçlarının, DÖC-BKD kuralının sonuçlarından belirgin bir şekilde daha iyi olduğu görülmüştür (Halıcı 2000-b). Öğrenme fazında, ÖB-BKD kuralı ile öğrenen sistem, başarılı eyleme daha iyi yakınsamaktadır. Sönüm fazında ise, ÖB-BKD ile öğrenen sistemde sönüm istendiği biçimde oluşmakta ve sistem çevresel değişimlere iyi bir uyum gösterebilmektedir. Çok adımlı kararlar için daha sonra yaptığımız benzetim çalışmaları da ÖB-BKD kuralı ile öğrenen sistemin DÖC-BKD ile öğrenen sisteme göre öğrenme ve sönüm açısından çok daha başarılı olduğunu göstermiştir.

## KAYNAKLAR


Carlson NR. Physiology of Behaviour, Allyn and Bacon-1977.

Gelenbe E. andom neural networks with positive and negative signals and product form solution. Neural Computation-1989; 1: 502-510.

Gelenbe, E. Stability of the random neural network model. Neural Computation-1990; 2: 239-247.

Gelenbe E. Learning in the recurrent random neural network. Neural Computation-1993; 5: 154-164.

Gelenbe E. Special issue on G networks, European J. of Operational Research-2000; 126.

Halıcı, U. Reinforcement learning in random neural networks for cascaded decisions. J. of Biosystems-1997; 40: 83-91.

Halıcı U. Biyolojik nörondan yapay sinir ağına. In: Beyin ve Kognisyon, edt by Karakaş S, Aydın H, Özesmi Ç, Özdemir C, Çizgi Tıp Yayıncılık-2000; 37-49.

Halıcı, U. Reinforcement learning with internal expectation for the random neural networks. European J. of Operational Research-2000; 126: 288-307.

Hulse HS, Egeth H, Deese J. he Psychology of Learning, McGraw Hill-1980.

Narendra K, Thathachar MAL. Learning Automata: An Introduction, Prentice Hall-1989.