ÖDEV 4

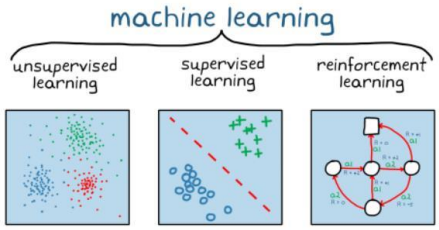
Selen Erdoğan
Selenerdogan2019@gtu.edu.tr
Elektronik Mühendisliği Bölümü, GTÜ

I. GİRİŞ

Ödevde yapay zekanın alt dallarından biri olan pekiştirmeli öğrenme (reinforcement learning) kullanılarak, belirli bir ızgarada başlangıç noktasından hedef noktasına en iyi yolun nasıl bulunabileceği ele alınmaktadır.Q-learning algoritması kullanılarak, bir ajanın karşılaştığı her durumda hangi hareketin en iyi sonucu vereceğini öğrenmesi hedeflenmektedir.

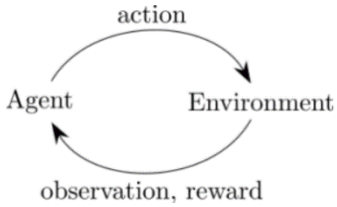
A. Pekiştirmeli Öğrenme

Pekiştirmeli öğrenme (reinforcement learning—RL), makine öğrenimi alanının önemli bir dalıdır. Bu yaklaşım, insanların ve hayvanların deneyimlerinden öğrenme süreçlerine benzer şekilde, algoritmaların ve yapay zekâ sistemlerinin ödül ve ceza mekanizmalarını kullanarak öğrenmelerini sağlar. Pekiştirmeli öğrenme algoritmaları, önceden etiketlenmiş verilere dayalı değil, ancak doğrudan etkileşimlerle ve denemelerle öğrenirler.



Şekil 1

Pekiştirmeli öğrenme; ajan (agent) adı verilen öğrenen bir sistem, durum (state), eylem (action) ve ödül (reward) unsurlarının etkileşimi üzerine kuruludur. Ajan, belirli bir durumda alabileceği eylemleri değerlendirerek en yüksek toplam ödülü elde etmeyi amaçlar. Bu süreçte ajan, deneme-yanılma yöntemiyle hareket eder ve aldığı ödüllerle performansını geliştirir.



Şekil 2

B. Q-Learning

Q-öğrenme, ajanın deneyimleri üzerinden öğrenme sağlayan ve her bir durum için en iyi eylemi belirlemeye çalışan model-temelli olmayan bir pekiştirmeli öğrenme algoritmasıdır. Ajan, karşılaştığı her durum için belirli bir ödül (veya ceza) alır ve bu ödüller, Q-değerleri tablosunda saklanır. Bu değerler, ajanın hangi durumda hangi hareketi yapması gerektiğini belirler.

II. ALGORİTMANIN OLUŞTURULMASI

Her bir durum (grid'in her bir hücresi) için bir Q-değer tablosu oluşturuldu. Bu tablo, ajanın her bir durumda alabileceği her bir eylem için beklenen ödülü (Q-değeri) tutmaktadır. Ajan, grid üzerinde rastgele veya bir politikaya bağlı olarak hareket etmesi sağlandı. Başlangıçta bu rastgele eylemlerle başlatıldı. Ajan her eylem yaptığında, algoritma ödülü alır ve Q-değerini günceller. Q-değer güncellemesi, Şekil 3'te görülen formülle yapıldı.

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

Burada s mevcut durumu, a mevcut eylemi, r alınan ödülü, s' yeni durumu ve a' yeni durumda alınabilecek eylemleri temsil eder.

Şekil 3

Bu işlem, ajan T noktasına ulaşana kadar veya belirli bir iterasyon sayısına ulaşana kadar tekrarlanır. Eğitim tamamlandıktan sonra, her bir durum için en yüksek Q-değeri olan eyleme karşılık gelen bir politika oluşturuldu.

III. ALGORİTMANIN UYGULANMASI

Algoritmada, belirlenen bir hedefe ulaşma amacı güden bir ajanın ızgara üzerindeki hareketlerini optimize etmeyi amaçlandı. Bu süreçte, ajanın karşılaştığı her bir durum ve aldığı her bir eylem, Q-tablosu adı verilen bir değer matrisi içerisinde kayıt altına alındı. Ajan, ızgaranın başlangıç noktasından yola çıkarak, hedefe ulaşıncaya kadar veya belirlenen bölüm sayısına ulaşıncaya kadar denemeler yaptı. Her deneme, ajanın çevresini ve alabileceği ödülleri keşfetmesi sürecidir ve bu, onun keşif stratejisini oluşturur. Keşif sırasında ajan, bazen düşük Q-değerlerine sahip eylemleri bile seçebilir; bu, daha önce az keşfedilmiş olan yolları denemesine olanak tanır.

Diğer yandan, ajanın öğrenme süreci ilerledikçe, Q-tablosunda biriken deneyimlerden yararlanarak daha bilinçli kararlar alması beklendi. Bu, sömürü stratejisine işaret eder ve ajan, en yüksek Q-değerine sahip eylemleri seçmeye başladı. Böylece, daha yüksek ödüllere yol açan hareketler tercih edildi. Her bölümde ajan, aldığı ödülü

maksimize etmek için Q-tablosunda biriktirilen bilgiye dayanarak hareket etti. Q-değeri, mevcut durum ve seçilen eylemin bir fonksiyonu olarak hesaplandı ve gelecekteki durumlar için beklenen toplam ödülü temsil edildi. Ajan bu şekilde, hem kısa vadeli hem de uzun vadeli ödüller arasında denge kurmayı öğrenir.

Ödül, ajanın hedefe ulaştığında 1.0, diğer durumlarda ise 0.0 olarak tanımlandı. Ajan, hedefe ulaştığında maksimum ödülü aldı ve bu, Q-tablosunun ilgili kısmını güncellemek için bir sinyal görevi görür. Q-değerleri, alınan ödülün yanı sıra gelecekteki durumlar için beklenen en yüksek Q-değerini de göz önünde bulundurarak güncellendi.

IV. SONUÇLAR

```
Bölüm: 998, Durum: (2, 0), Hareket: down, Ödül: 0
Bölüm: 998, Durum: (1, 0), Hareket: up, Ödül: 0
Bölüm: 998, Durum: (0, 0), Hareket: up, Ödül: 0
Bölüm: 998, Durum: (0, 1), Hareket: right, Ödül: 0
Bölüm: 998, Durum: (0, 2), Hareket: right, Ödül: 1
Bölüm: 999, Durum: (1, 0), Hareket: up, Ödül: 0
Bölüm: 999, Durum: (0, 0), Hareket: up, Ödül: 0
Bölüm: 999, Durum: (0, 1), Hareket: right, Ödül: 0
Bölüm: 999, Durum: (0, 2), Hareket: right, Ödül: 1
Bölüm: 1000, Durum: (1, 0), Hareket: up, Ödül: 0
Bölüm: 1000, Durum: (0, 0), Hareket: up, Ödül: 0
Bölüm: 1000, Durum: (0, 1), Hareket: right, Ödül: 0
Bölüm: 1000, Durum: (0, 2), Hareket: right, Ödül: 1
```

Şekil 4

Şekil 4'de bölümler (episodes), ajanın eğitim sürecinin her bir adımını temsil eder. Her bölümde ajanın durumu, aldığı hareket ve bu hareketin sonucunda aldığı ödül gösterilmektedir. Ajanın yaptığı hareketler ve aldığı ödüller, gelecekteki kararlarını etkileyecek Q-değerlerinin güncellenmesi için kullanıldı.

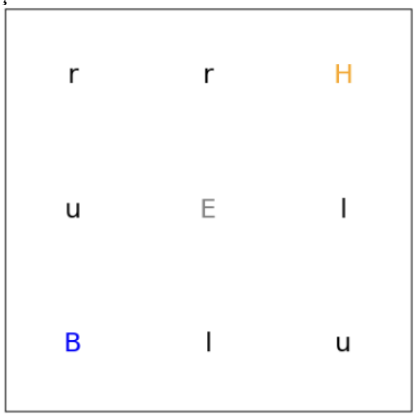
```
Öğrenilen politika:
right right Hedef
up Engel left
Başla left up

Q-tablosu değerleri:
DURUM (0, 0), Q DEĞERLERİ: [0.81 0.9  0.73 0.81]
DURUM (0, 1), Q DEĞERLERİ: [0.9  1.   0.66 0.81]
DURUM (0, 2), Q DEĞERLERİ: [0. 0. 0. 0.]
DURUM (1, 0), Q DEĞERLERİ: [0.81 0.66 0.66 0.73]
DURUM (1, 1), Q DEĞERLERİ: [0.68 0.   0.3  0.73]
DURUM (1, 2), Q DEĞERLERİ: [0.   0.   0.   0.33]
DURUM (2, 0), Q DEĞERLERİ: [0.73 0.59 0.66 0.66]
DURUM (2, 1), Q DEĞERLERİ: [0.   0.   0.   0.66]
DURUM (2, 2), Q DEĞERLERİ: [0. 0. 0. 0.]
```

Şekil 5

Şekil 5 çıktısında görülen "öğrenilen politika" kısmı, ajanın her bir durum için öğrendiği en iyi hareketi gösterir. Bu politika, ajanın eğitim süreci sonucunda ulaştığı ve hedefe ulaşmak için izlemesi gereken yolu belirler.
"Q-tablosu değerleri" kısmı, her bir durum için ajanın tahmin ettiği uzun vadeli ödüllerin sayısal değerlerini içerir. Bu değerler, ajanın her durumda hangi hareketi yapması gerektiğine karar verirken kullanılır ve en yüksek değere sahip hareket genellikle tercih edilir.

Bu döngünün, ajanın hedefe ulaşmasını sağlayacak stratejiyi bulana kadar veya belirli bir iterasyon sayısını tamamlayana kadar devam etmesi sağlandı. Sonuçta, ajanın her bir durum için en uygun eylemi seçebileceği, yani hedefe en verimli yoldan ulaşabileceği bir politika ortaya çıktı.



Şekil 6

Q-learning algoritması, özellikle başlangıç aşamasında, eylem seçiminde belirli bir miktar rastgelelik (epsilon-greedy yöntemi) kullanır. Bu nedenle, her çalıştırmada farklı sonuçlar üretebilir. Eğer epsilon değeri yeterince azalmazsa veya yeterli sayıda bölüm (episode) çalıştırılmazsa, algoritmanın öğrenmesi tamamlanmayabilir ve farklı çalıştırmalarda farklı politikalar öğrenebilir.
Q-tablosunun başlangıç değerleri algoritmanın öğrenme sürecini etkileyebilir. Eğer bu değerler rastgele olarak belirlenmişse veya her çalıştırmada farklı başlangıç değerleri kullanılmışsa, bu durum farklı politikaların öğrenilmesine yol açabilir.
Öğrenme oranı (alpha) ve indirim faktörü (gamma) gibi parametrelerin değerleri, algoritmanın karar verme sürecini ve sonuçta elde edilen politikayı etkiler. Bu parametrelerin her çalıştırmada aynı olup olmadığını kontrol etmek önemlidir.
Q-learning algoritmasında, güncellenen durumların sırası ve seçilen eylemler sonucu öğrenilen politikayı etkiler. Eğer algoritma her bölümde farklı durumlardan başlıyor veya farklı eylemler seçiyorsa, bu da farklı sonuçlar üretebilir.

KAYNAKÇA

https://medium.com/@vlknlvnt/reinforcement-learning-peki%C5%9Ftirmeli-%C3%B6%C4%9Frenme-i%CC%87nsan-beyniyle-aradaki-fark-kapan%C4%B1rken-c0d0a6f7c2e8
https://miuul.com/not-defteri/pekistirmeli-ogrenmeye-k%C4%B1sa-bir-giris