Çok Etmenli Pekiştirmeli Öğrenmede Devingen Ortamlarda Bağlam Değişim Tespiti ve Tanımlama
Context Detection and Identification In Multi-Agent Reinforcement Learning With Non-Stationary Environment

Ekrem Talha SELAMET, Borahan TÜMER
Bilgisayar Mühendisliği Bölümü
Marmara Üniversitesi
İstanbul, Türkiye
ekremtalhaselamet@gmail.com, borahan.tumer@marmara.edu.tr

2022 30th Signal Processing and Communications Applications Conference (SIU) | 978-1-6654-5092-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SIU55565.2022.9864802

***Özetçe* —Pekiştirmeli öğrenme yaklaşımları, çoğunlukla ortamın durağan olması varsayımıyla etmenin öğrenmesini konu alır. Fakat, gerçek hayat uygulamalarında ortam durağan değildir. Birçok durağan ortamın bir araya gelmesiyle oluşan devingen ortamlardır. Ortamda birden fazla etmen bulunabilir ve bu etmenler de ortamı devingen hale getirmektedir.** *Pekiştirmeli öğrenme-bağlam sezme* **(RL-CD) [1] yöntemi, etmenin devingen ortam hakkında önsel bir bilgisi olmadan öğrenmesini ve bağlam değişimlerinin belirlenmesini sağlayan yaklaşımdır. Bu yaklaşımın temelindeki ortamda tek etmen vardır ve çok etmenli öğrenim için eksiklikleri bulunmaktadır. Bu çalışmada çok etmenli devingen ortamlarda hem bağlam değişim noktalarını sezebilen hem de etmenlerin ortamları öğrenebilmesine olanak sağlayan** *çok etmenli pekiştirmeli öğrenme-bağlam sezme* **(MARL-CD) adında yeni bir yaklaşım geliştirilmiştir. Bu yaklaşım RL-CD yöntemini temel alır. Çok etmenli öğrenmede, etmenlerin ortam üzerinde oluşturdukları devingenliği sezmesi ve bağlam değişikliğini belirlemesi yönüyle daha verimlidir. Bağlamdaki değişiklikleri yalnızca ortam dinamiklerinin değişiminin yanı sıra ortamdaki etmenlerin politika değişikliklikleriyle de belirleyebilmesini sağlar. Bu çalışmadaki yaklaşımda, etmenler enerjilerini %16 daha az harcayarak ve değişim noktalarını daha doğru sezmesi açısından RL-CD'ye daha verimli olduğu, deney sonuçları ile gösterilmiştir.**

Anahtar Kelimeler—Pekiştirmeli öğrenme, devingen ortamlar, bağlam sezme, çoklu etmenli öğrenme.

***Abstract*—Reinforcement learning methods are mostly constructed on the very assumption that environments are stationary. However, most real world environments are non-stationary; that is, we assume they are composed of several stationary components (i.e., sub-environments or** *contexts*) **. So, methods with this assumption are not capable of learning non-stationary environments.** *Reinforcement Learning - Context Detection* **(RL-CD) method enables the agent to learn the environment without prior information; detect the environment's context change points and create a partial model for each context. The underlying environment of this approach is single-agent and has shortcomings for multi-agent learning. In this study, we introduce a new approach called** *Multi-agent reinforcement learning-context detection* **(MARL-CD), which can both detect context change points and enable agents to learn non-stationary environments with multi-agent settings. This approach is based on RL-CD approach. MARL-CD is more efficient in terms of detecting context change created by the agents on the environment and detecting the context change of the environment itself. It enables an agent to detect the context changes not only from the change of environment dynamics but also from policy changes of agents in the environment. In the approach in this study, it has been shown by the experimental results that the agents spend 16% less energy and are more efficient than RL-CD in terms of detecting the change points more accurately.**

Keywords—Reinforcement learning, non-stationary environment, context detection, multi-agent

I. GİRİŞ

Pekiştirmeli öğrenme (PÖ), davranışsal öğrenmenin bir çeşidi olup, bir etmenin ortam ile sürekli etkileşimler kurup sonucunda pekiştirme (ödül ya da ceza) alarak öğrenmesini hedefler. Etmenin esas amacı, gerçekleştirdiği eylemler sonucu gözlemlerini, ortamın verdiği ödül veya cezayı ve bulunduğu durumu, kullanarak ödüllerin toplamını en üst düzeye getirmesini sağlayacak çözümü bulmaktır.

Klasik PÖ yöntemleri ortamların durağan olduğu varsayımı üzerine kuruludur. Ancak, gerçek ve doğal ortamların çoğu devingendir; diğer bir deyişle ortamlar birkaç farklı durağan bağlamdan oluşmaktadır. Bundan dolayı, bu ortamın durağan olduğu varsayımına sahip yöntemler bu ortamları öğrenmekte yetersiz kalmaktadır. Bu yöntemler o andaki ortam bağlamını öğrenirler, fakat ortam bağlam değiştirdiğinde eski bağlam unutulur ve yeni ortam bağlamını öğrenme süreci başlar. Dolayısıyla etmen geçmiş deneyimlerini, yeni bağlamı öğrenirken unutmaktadır. Klasik PÖ yöntemleri, devingen ortamları öğrenememektedir.

Çalışmamız için önemli bir çıkış noktası olan *pekiştirmeli öğrenme-bağlam sezme* (RL-CD) yöntemi [1], etmenin ortamdaki değişimi sezmesini ve değişken ortamın her bir durağan bağlamına ait parçalı modeller oluşturmasını sağlar.

978-1-6654-5092-8/22/$31.00 ©2022 IEEE

Böylelikle etmen, geçmişe ait deneyimlerini yitirmezken, aynı ortam durumlarıyla yeniden karşılaştığında öğrenme sürecini tekrarlamadan ilgili parçalı modelde öğrenmiş olduğu çözümü anımsar ve daha hızlı öğrenir. Etmenin belleğindeki her bir bağlama ait parçalı modelin, ortamın o andaki bağlamıyla uyum niteliği, ortamın o andaki bağlamını ne kadar iyi yansıttığına göre hesaplanır. Model uyum niteliği için belirlenecek olan eşik değer, ortamın o andaki bağlamının belirlenmesi için kullanılır. Eğer etmenin belleğindeki parçalı modeller eşik değerin altındaysa, hiçbir model o andaki ortam bağlamını yeterince güçlü temsil etmediğinden yeni bir parçalı model oluşturulur.

Gerçek hayat uygulamalarında ortamlar çoğunlukla çok etmenli sistemlerden (ÇES) oluşur. Bu sistemlerde birçok etmen ortamla etkileşimler kurarak öğrenmeye çalışır. Bu etkileşim esnasında ortamın dinamik değişkenlerine ek olarak etmenlerden kaynaklanan devingenlik de meydana gelir. Örneğin, etmenlerden biri her zamanki politikasının dışına çıkarak diğer etmenleri etkileyecek seçimler yapabilir. Diğer etmenlerin bunu fark ederek politikalarını buna göre uyarlaması gerekir. RL-CD tek etmenle çalışan bir yöntemdir. Bundan dolayı çok etmenli ortamlarda diğer etmenlerin hareketlerine göre kendisini adapte etmekte ve etmenin içinde bulunduğu bağlam, ortam kaynaklı değişmediğinde bunu sezme konusunda yetersiz kalmaktadır.

Bu çalışmada (MARL-CD), RL-CD yöntemi temel alınarak çok etmenli devingen ortamlarda bağlam değişiminin sezilmesi ve tanımlanması için yeni bir yöntem geliştirilmiştir. Bu yöntemde her etmen ortamda bulunan diğer etmenleri gözlemleyerek her birisi için RL-CD yönteminde önerilen parçalı modelleri oluşturur. Böylelikle sadece kendi modeli üzerinden değil, gözlemlediği diğer etmenlerin modelini kullanarak ortam bağlamındaki değişimleri belirleyebilmektedir. Ayrıca, etmen gözlemlediği diğer etmenlerin modellerini kullanarak, seçeceği eylemlerde daha önce keşfetmediği bir ortam bölgesindeyse ya da diğer etmenlerin eylemleri sonucu aldığı ortam yanıtlarını da göz önüne alarak seçimler yapabilir. Diğer etmenleri gözlemlemesinin sonucu olarak, ödüllerin toplamını en üst düzeye getirmesini sağlayacak diğer politikaları da kullanabilmesine olanak sağlar. MARL-CD, çok etmenli devingen ortamlarda bağlam değişiminin sezilmesine ve tanımlanmasına olanak sağlaması, etmenin diğer etmenlerin politika değişikliklerini sezebilmesini sağlaması ve eylem seçim kararlarını sadece kendi modeline bağlı kalmadan alabilmesi açılarından RL-CD yönteminden ayrılmaktadır.

MARL-CD'nin çeşitli açılardan daha verimli olduğu deneyler ile gösterilmiştir. Deneylerde ortamlar birçok durağan ortamın birleşiminden oluşan devingen ortamlardır. Ayrıca her etmen eşit yeteneklerde olup, ortamın verdiği tepkiler aynı durum için her etmene eşittir. Bu koşullar altında sonuçlar incelendiğinde; etmenlerin maksimize ettiği toplam ödülün RL-CD'den daha yüksek olduğu ve ortam üzerinde hareket ederlerken harcadıkları toplam enerjinin de minimize edildiği gözlenmiştir.

Bu bildirinin geri kalanı şu şekilde organize edilmiştir. İkinci bölümde ilgili geçmiş çalışmalar ve MARL-CD'den farklılıkları hakkında bilgi verilmiştir. MARL-CD'nin nasıl temellendirildiğine dair kuramsal bilgi ve bu yaklaşımın ayrıntıları üçüncü bölümde açıklanmıştır. Dördüncü bölümde deney sonuçları yorumlanmıştır. Son bölümde ise bu çalışmanın sonucunda elde edilen bilgiler özetlenmiştir.

II. İLGİLİ ÇALIŞMALAR

Gerçek ve doğal ortamların devingen (zaman içinde değişen karakterde) olmaları (örneğin ardışık oluşan farklı durağan bağlamların oluşturduğu bir ortam) PÖ alanı için zorlayıcı bir faktördür. Diğer bir deyişle devingen ortamlar birçok durağan bağlamın birleşmesinden oluşur. Q-Learning [3], Prioritized Sweeping (PS) [2] gibi klasik PÖ yöntemleri ortamların durağan olduğu varsayımına dayanır. Bu yaklaşımlarda etmen ortamı öğrenir. Ortam değiştiğinde ise, bağlam değişikliğini sezemez ve yeni bağlamı öğrenmeye çalışır. Bu sırada etmen geçmiş deneyimlerini unutur. Sonraki paragraflarda, bahsedilen unutma probleminin çözümü için bu çalışmayla da benzerliği bulunan önerilmiş diğer yaklaşımlardan bahsedilecektir.

Choi et. al. [4] devingen ortamların çalışma kiplerinden oluştuğunu ve bu kiplerin devingen ortamın durağan parçaları olduğu fikrini sunmuştur. Bu yaklaşımla birlikte, devingen ortamlar gizli-kip Markov karar süreçleri (HM-MDP) haline gelmiştir. Kipler farklı MDP'lerdir ve gizli kip modelleri ise sonlu MDP'lerdir. MDP'ler aynı durum ve eylem kümesine sahip farklı ödül ve geçiş fonksiyonlarıdır. Önerilen yeni süreç, gizli-kip Markov karar süreçleri, bir MDP tarafından durum geçişleri ile denetlenebilir. HM-MDP'lerde durumlar gözlemlenebilirken kipler gözlemlenemez. Gizli-kip modeli, Baum-Welch algoritmasının varyasyonu ile öğrenilir.

Da Silva et. al. [1] RL-CD adında bir yaklaşım önermiştir. Bu çalışmada da ortamın farklı durağan *bağlam*lardan oluşan devingen bir ortam olduğu varsayılmıştır. Bu bağlamlar, etmenin ortamla etkileşimleri sonucu gözlemlediği durum geçişleri ve aldığı ödüller olarak temsil edilir. Bu bağlam belirleme yöntemiyle birlikte etmen, bağlam değişim noktalarını sezer ve ortamın parçalı modellerini yaratır. Böylelikle, etmen bağlamları tanır. Etmen, bağlamı parçalı modelini yaratarak öğrenir. Bağlam değişiminden sonra etmen, yeni karşılaşılan bağlamı, daha önce yarattığı parçalı modellerle, etmenin politikası sonucunda oluşan ve politikanın ilgili parçalı bağlam modeliyle uyumunu tanımlayan kalite sinyaliyle karşılaştırır. Kalite sinyali, parçalı modelin o bağlama ne kadar uygun olduğunu gösterir. Eğer etmen daha önce karşılaştığı bir bağlamı tekrar sezerse, ilgili parçalı modelini tekrar kullanabilir. Eğer etmen yeni bir bağlam ile karşılaşırsa yeni bir parçalı model yaratır. Bu yaklaşımda etmen aynı bağlamı tekrar öğrenmek zorunda kalmaz ve tekrar öğrenme süreci yerine geçmiş deneyimleriyle oluşturduğu parçalı modeli kullanarak öğrenmeyi hızlandırır.

Devingen ortamlarda bağlamlar arası geçişi belirlemede rastlantısal öğrenmeyle zayıf değerlemeye (SLWE) dayalı bağlam değişimlerinin sezilmesi [9] yöntemi de geliştirilmiştir. SLWE [10] yöntemini kullanarak birinci dereceden markov (FOM) olasılıklarını bağlam değişim noktalarını belirlemek için kullanır. Yeni bağlam için ortamın ürettiği dizi ögelerinin FOM bağımlılığını yansıtan yeni bir değere yeniden yakınsayarak öğrenir ve değişimleri sezer.

Çok etmenli ortamlarda, tek etmenli öğrenme yaklaşımları yetersiz kalmaktadır. Ortamdaki diğer etmenlerin politika değişikliklerini sezememeleri ve bu değişikliklere uyum sağlayıp diğer etmenlere yanıt verebilecek farklı politikalar geliştiremediklerinden, tek etmenli yaklaşımlar eksik kalmaktadır. Bu

eksikliklerden dolayı, ayrıca, zaman içerisinde öğrendikleri en iyi politikalar, en iyi olmayan politikalara dönüşebilir. Rakip etmenlerin politikalarını izlemek rekabetçi çok etmenli ortamlar için en verimli yöntemdir. *Bayes-XCS* [6] yöntemi, Markov oyunlarında XCS [7] yöntemini temel alan, rakip etmenin davranışlarını kullanarak en iyi cevabı öğrenir. Rakip etmen tarafından üretilen veriyi toplayarak buna karşılık gelen rakip modelleri oluşturulur. Bu modeller kullanılarak rakip etmenin politika değişiklikleri belirlenir. Ayrıca Bayezyen Politika Yineleme (BPR) yöntemini de [5] temel alarak geliştirilen Bayes-XCS yöntemi rakip etmenlerin modellerini kullanarak etmenin en iyi yanıt politikasını tekrar kullanmasına yardımcı olur.

III. METODOLOJİ

Bu çalışmada RL-CD yaklaşımını temel alan çok etmenli devingen ortamlarda öğrenme yaklaşımı sunulmuştur. MARL-CD, aynı ortamdaki etmenler ortam ile etkleşimleri sırasında pekiştirme alarak öğrenmelerinin yanı sıra ortamdaki diğer etmenleri de izleme yetisine sahiptir. Böylelikle ortamdaki rakip etmenlerin gözlemledikleri durum ve eylem bilgilerini kullanarak o etmene ait parçalı modelleri oluşturabilmektedir. Etmen kendi öğrenmesini RL-CD yöntemini kullanarak gerçekleştirirken, gözlemlediği etmenin modelini de RL-CD yöntemi kullanarak oluşturur. Bu çalışma üç yönüyle RL-CD yönteminden ayrılmaktadır: i) etmenin ortamdaki diğer etmenleri izleyerek parçalı modellerini oluşturması; ii) etmen, ortamdaki diğer izlediği etmenlerin eylem ve durum izlerini çözümleyerek rakip etmenlerin politikalarını kendi politikalarıyla karşılaştırarak daha iyi politikalar yaratmasına olanak sağlaması; iii) etmenlerin aynı birim zaman içerisinde ardışık çoklu eylem seçebilme *becerisine* olanak sağlaması.

MARL-CD ile birlikte etmen yalnızca kendi modeline bakarak bağlam değişim noktalarını sezmez. İzlediği diğer etmenlerin modellerini kullanarak da bunu yapabilir. Ortamın devingenliğini yalnızca ortamın kendi dinamikleri (farklı bağlamlar arası geçişler) değil, rakip etmenlerin politika geçişleri de sağlamaktadır. Etmen yalnızca kendi modelini dikkate aldığında, sezdiği ortam değişikliğinin bağlam değişikliğinden mi yoksa izlediği etmenin politika değişikliğinden mi kaynaklandığını belirleyemez. Oysa MARL-CD'de rakip etmenleri izlediği için, hangi parçalı modeli kullanacağına daha doğru şekilde karar verir ve politikasını buna göre değiştirebilir. Ayrıca ortam yalnızca küçük değişikliklerle bağlam değiştirdiğinde, model kalite sinyalindeki değişim değişim noktası sezme eşik değerinin altında kalabilir. Bu durumda da; birden fazla bağlam modeline bakarak karar almak küçük bağlam değişikliklerini de yakalamasını sağlar. Etmenler keşifsel seçimler sonucu ortamın farklı durumlarını deneyimleyebilirler. Böylece etmen ilk defa bulunduğu durumda seçimler yaparken eğer kendi modelinde bu durum yoksa izlediği etmenlerin modeline bakarak durum-ödül değerlerini dikkate alacak seçimler yapabilir. Böylelikle etmen keşifsel seçimler yapmadan da durum hakkında bilgi sahibi olabilir. Öğrenme hızını artırabilir.

Etmen, ortamdaki rakip diğer etmenlerin eylem ve durumunu izlediğinden dolayı eylem ve durum izlerini çözümleyebilir. Eğer izlediği etmen kendisinden daha başarılı bir politika izliyorsa bu politikayı geçici bir süreliğine benimseyebilir. Etmen kendi modeli üzerinde bağlam değişim noktası (BDN) sezerse rakip etmenlerin politika değişimi yapıp yapmadığını da eylem ve durum izlerini çözümleyerek belirleyebilir. Böylelikle bağlam değişiminin kaynağına göre parçalı model seçimini gerçekleştirebilir.

Beceri (options) [8],temel eylemlerin genişletilmesinden oluşan eylemler dizisidir. Bir *Markov beceri*si'nde öncelikle bir sonraki eylem, politikaya bağlı olarak geçiş olasıklıklarına göre seçilir. Belirli bir olasılıkla beceriyi sonlandıracak bir sonraki duruma ortam tarafından geçiş gerçekleştirilir. MARL-CD'de tanımlanan bu beceri kullanılmaktadır. Etmen bulunduğu duruma bağlı olarak becerisini seçebilmektedir. Aynı birim zamanda birden fazla durum değişikliği yapma yetisi kazanmaktadır. Böylece etmenler hızlarını değiştirebilir. Fakat bunun karşılığında etmenin enerjisinde, tek bir eylem seçmesi sırasında harcayacağı enerjiden daha fazlası harcanmış olur. Bu durum ise etmenin alacağı ödülü etkiler. Böylelikle etmenler birbirlerine üstünlük kurmak amacıyla ya da rakip etmenleri şaşırtmak amacıyla becerileri kullanabilir. Bunun yanında; etmenin birim zamanda daha çok durum geçişi yapmasını sağlayabileceğinden, bu beceri öğrenmeyi hızlandırır.

MARL-CD, yukarıda bahsedilen uygulamalar bağlamında RL-CD ile karşılaştırılarak; çok etmenli devingen ortamlarda, etmenlerin enerjilerini daha verimli kullanmaları, bulundukları bağlamları daha hızlı öğrenmeleri ve bağlam değişim noktalarını daha hızlı sezmeleri açısından RL-CD yönteminden daha verimli olduğu deneylerle gösterilmiştir.

IV. DENEY SONUÇLARI

A. Ortam

Önerilen yaklaşımın sonuçları deneyler ile doğrulanmıştır. Deneylerde RL-CD'nin de gerçeklendiği top yakalama örneği (ball catching) [1] kullanılmıştır. Ortam ayrık toroidal 15x15 hücreden oluşan grid sistemdir. Etmen kedi, hedef ise hareketli topun yakalanmasıdır. Top rastgele bu grid sistem üzerinde herhangi bir yerde hareketine başlar. Topun eylem kümesi 4 ana yöndür. Etmenin hareket kümesi ise 4 ana yöne ek olarak hareket etmeme seçeneğiyle birlikte beş farklı eylemden oluşmaktadır. Bu ortamın devingenliği topun hareketlerindeki değişimden kaynaklanmaktadır. Top zaman içerisinde belirli bir yönde hareket ederken diğer yönde harekete geçerek davranış değiştirmektedir. Etmenler, ortamda aynı durumdan başlamaktadırlar ve aynı durumda aynı anda birden fazla etmen bulunabilir. Etmenler, seçtikleri hareketler ile diğer etmenleri doğrudan etkilememektedir. Bölümü kazanan etmen ödül alırken kaybeden etmenler aynı oranda ceza almaktadır ve bir sonraki bölüm ile devam edilir. Bu ortamda hem RL-CD yaklaşımı hem de MARL-CD test edilip sonuçları karşılaştırılmıştır.

B. Yöntemler

Deneyler aynı koşullar altında iki farklı yaklaşım içinde gerçeklenmiştir. Birinci kurulumda etmenler RL-CD ile öğrenirken, ikinci kurulumda etmenler MARL-CD yöntemi ile öğrenmiştir. Deney aşamasında farklı parametreler ile deneyler gerçeklenmiştir. İki yaklaşımın oyun kazanma sayıları ve bağlam değişim noktası belirleme sayıları karşılaştırılmıştır. Bunun yanında etmenlerin deney süreci boyunca enerji değişimleri de karşılaştırılmıştır.

C. Sonuçlar

Deneyler sonucunda Tablo I'de görüleceği üzere RL-CD ile öğrenen etmenler, BDN'yi belirleme konusunda yetersiz

kalmaktadır. BDN sezme sayısının yüksek olması bunu göstermektedir. Ortam bağlamı değişmese bile rakip etmenin politika değişikliklerini ortam bağlam değişimi olarak algılamaktadır. Bu da etmenin bulunduğu bağlamı öğrenmesini zorlaştırırken aynı zamanda yanlış bağlama ilişkin parçalı modelleri kullanması nedeniyle parçalı modellerin kalitesini de düşürmektedir. Etmenlerin kazandığı oyun sayıları bağlam başına değerlendirildiğinde genellikle bir etmen diğerine üstünlük kurmaktadır.

MARL-CD sonucunda Tablo II'de görüleceği üzere etmenler bağlam değişim noktalarını doğru şekilde belirleyebilmiştir. RL-CD'nin aksine bağlama uygun olan parçalı modeller kullanılmıştır. Böylece etmenlerin yapmış olduğu seçimler, toplam ödül miktarını maksimize edecek şekilde yapılmıştır. Ayrıca etmenlerin kazandığı oyun sayıları RL-CD ile karşılaştırıldığında bir etmenin diğerine tamamen üstünlük kurmadığı, başarımlarının beklendiği gibi daha dengeli olduğu görülmüştür.

Etmenlerin bağlam başına harcadığı enerji miktarlarına baktığımızda, MARL-CD kullanan etmenlerin ortalama %16 daha az enerji harcadığı görüldü. Bu fark Şekil 1'de görülmektedir. Şekil 1 üzerindeki dikey kesikli çizgiler bağlam değişim noktalarını temsil etmektedir. Daha verimli olmasında MARL-CD'nin daha hızlı öğrenmesi, doğru parçalı modele daha hızlı karar vermesinden dolayı daha az hareket etmesi etkilidir. Ayrıca MARL-CD, etmenler becerikli seçimler yapabilmesinden dolayı öğrenme süresi de kısalmaktadır.

TABLO I: RL-CD OYUN KAZANMA VE BAĞLAM DEĞİŞİM NOKTASI SEZME SONUÇLARI

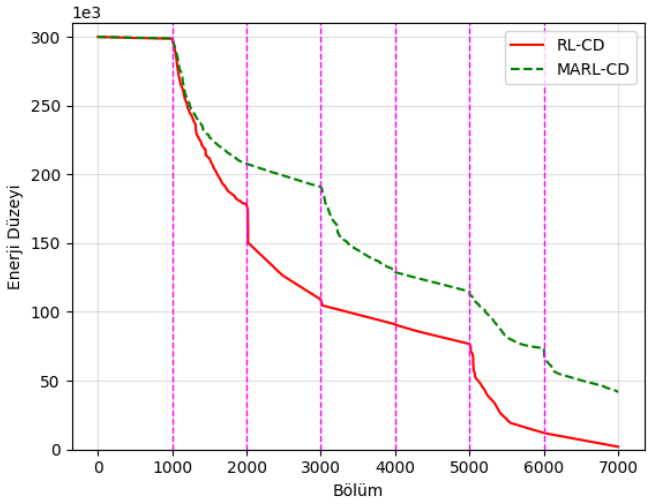
	Oyun Kazanma Sayıları		Bağlam Değişim Noktası Sezme Sayıları	
Bağlam	Etmen 1	Etmen 2	Etmen 1	Etmen 2
0	917	721	1	1
2	566	438	1111	2521
0	981	21	755	37
1	990	89	2	4
0	714	287	1475	33
2	697	314	3	48
1	1000	2	2	1

V. ÇIKARIMLAR

Bu çalışmada çok etmenli devingen sistemler için yeni bir yaklaşım sunulmuştur. RL-CD yöntemini temel alan bir yaklaşım olmasına karşın; ortamdaki diğer etmenleri izleyerek parçalı modellerini oluşturması; etmenin, ortamdaki diğer izlediği etmenlerin eylem ve durum izlerini çözümleyerek rakip etmenlerin politikalarını kendi politikalarıyla karşılaştırarak daha başarılı politikalar yaratmasına olanak sağlaması; etmenlerin aynı birim zaman içerisinde arka arkaya eylem seçebilme (*becerilerinin*) bulunması yönüyle ayrılmaktadır. Deneyler sonucunda etmenlerin enerjilerini daha verimli kullanmaları, bağlam değişim noktalarını daha iyi sezebilmeleri açısından RL-CD yaklaşımına göre daha iyi başarım sergilediği görülmüştür.

TABLO II: MARL-CD OYUN KAZANMA VE BAĞLAM DEĞİŞİM NOKTASI SEZME SONUÇLARI

	Oyun Kazanma Sayıları		Bağlam Değişim Noktası Sezme Sayıları	
Bağlam	Etmen 1	Etmen 2	Etmen 1	Etmen 2
0	938	934	1	1
2	539	472	16	7
0	813	641	1	1
1	621	395	1	1
0	998	1000	1	1
2	692	543	1	1
1	499	802	1	1



Şekil 1: Etmenlerin algoritmalara göre enerji değişim grafiği

Önümüzdeki çalışmalarda, geliştirilen bu yöntemi Pac-Man ve gerçek hayata daha yakın ortamlarda deneyerek sonuçların incelenmesi planlanmaktadır. Bu çalışmaya sezgisel faktörlerin eklenmesi de planlanmaktadır. Etmenin, modelini kullanarak ortamın belirli durumlarında ödül alma olasılığı yüksek durumları sezgisel olarak tespit edebileceği bir metot daha eklenecektir. Böylece etmenlerin ortamı daha iyi ve hızlı öğrenerek enerji harcamasının azaltılması hedeflenmektedir.

KAYNAKLAR

[1] B. C. da Silva, E. W. Basso, A. L. C. Bazzan, and P. M. Engel, "Dealing with non-stationary environments using context detection," in Proceedings of the 23rd international conference on Machine learning - ICML '06, 2006.

[2] A. W. Moore and C. G. Atkeson, "Prioritized sweeping: Reinforcement learning with less data and less time," Mach. Learn., vol. 13, no. 1, pp. 103–130, 1993.

[3] C. J. C. H. Watkins and P. Dayan, "Q-learning," Mach. Learn., vol. 8, no. 3–4, pp. 279–292, 1992.

[4] S. P. M. Choi, D.-Y. Yeung, en N. L. Zhang, "An Environment Model for Nonstationary Reinforcement Learning", in Proceedings of the 12th International Conference on Neural Information Processing Systems, Denver, CO, 1999, bll 987–993.

[5] B. Rosman, M. Hawasly, and S. Ramamoorthy, "Bayesian policy reuse," Mach. Learn., vol. 104, no. 1, pp. 99–127, 2016.

[6] H. Chen, J. Huang, Q. Liu, C. Wang, and H. Deng, "Detecting and tracing multi-strategic agents with opponent modelling and Bayesian policy reuse," in 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2020.

[7] S. W. Wilson, "Classifier fitness based on accuracy," Evol. Comput., vol. 3, no. 2, pp. 149–175, 1995.

[8] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," Artif. Intell., vol. 112, no. 1–2, pp. 181–211, 1999.

[9] E. Aslanci, K. Coskun, P. Schuller, and B. Tumer, "Detection of regime switching points in non-stationary sequences using stochastic learning based weak estimation method," in 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), 2017.

[10] B Oommen and L Rueda. Stochastic learning-based weak estimation of multinomial random variables and its applications to pattern recognition innon-stationary environments. Pattern Recognition, 39(3):328–341, 2006.