한국경영과학회지
제40권 제3호
2015년 8월
http://dx.doi.org/10.7737/JKORMS.2015.40.3.063

폐쇄공간에서의 에이전트 행동 예측을 위한 MDP 모델

진효원[1] · 김수환[1†] · 정치정[2] · 이문걸[1]
[1]국방대학교 운영분석학과, [2]한국과학기술원 산업시스템공학과

MDP Modeling for the Prediction of Agent Movement in Limited Space

Hyowon Jin[1] · Suhwan Kim[1] · Chijung Jung[2] · Moongul Lee[1]
[1]Department of Operations Research, Korea National Defence University(KNDU),
[2]Department of Industrial and System Engineering, Korea Advanced Institute of Science and Technology(KAIST)

■ Abstract ■

This paper presents the issue that is predicting the movement of an agent in an enclosed space by using the MDP (Markov Decision Process). Recent researches on the optimal path finding are confined to derive the shortest path with the use of deterministic algorithm such as A* or Dijkstra. On the other hand, this study focuses in predicting the path that the agent chooses to escape the limited space as time passes, with the stochastic method. The MDP reward structure from GIS (Geographic Information System) data contributed this model to a feasible model. This model has been approved to have the high predictability after applied to the route of previous armed red guerilla.

Keywords : MDP, Optimal Path, Deterministic Algorithm, Stochastic, GIS

1. 서 론

임의의 대상(이하 에이전트)이 취할 행동이나 사건 발생 여부를 예측하는데 많이 사용되는 방법 중 하나는, 그 행동이나 사건이 랜덤(random)하게 발생한다고 가정하고 문제에 접근하는 것이다. 하지

논문접수일 : 2015년 05월 08일 논문게재확정일 : 2015년 06월 09일
논문수정일(1차 : 2015년 06월 02일, 2차 : 2015년 06월 03일)
† 교신저자, ksuhwan@kndu.ac.kr

만, '랜덤'이라는 가정은 에이전트가 특정한 목적 없이 무작위로 행동하거나, 에이전트의 행동 패턴을 사전에 파악하기 어려운 경우에 한해 적용하는 것이 바람직하다. 예를 들어, 전략과 전술에 근거해 활동하는 군사 집단이나 경찰의 추격을 피해 도주하는 범죄자의 행동 등을 예측해야 한다면, 랜덤 기반의 가정을 바탕으로 예측을 시도하는 것은 합리적이라고 볼 수 없다.

본 연구에서는 이처럼 계획적으로 행동하는 에이전트의 움직임을 예측하기 위해, 에이전트의 행동을 랜덤하게 보지 않고, 목적을 달성하기까지 보상(reward)을 가장 크게 받는 방식으로 행동할 것이라고 가정했다. 이 같은 가정을 바탕으로, 폐쇄된 공간에서 에이전트가 이동할 경로를 예측하는 효율적인 방법을 제시하고자 한다.

이동 경로(최적 경로)를 결정하는 방법에는 결정론적(deterministic)인 방법과 확률론적(stochastic)인 방법이 있는데, 본 연구에서는 확률론적 방법인 마코프 의사결정 프로세스(MDP; Markov Decision Process)를 활용, SSP(Stochastic Shortest Path) 관점으로 문제에 접근했다.

결정론적 방법의 최적 경로에 관한 연구로는 방수남의 연구가 있다[3]. 해당 연구에서는 지형정보(GIS)를 바탕으로 열상 감시장비의 탐지율을 추출, A*, Dijkstra 알고리즘을 사용해 탐지율이 낮은 지역을 연결하고 이를 에이전트의 이동 경로로 제시했다.

확률론적 방법의 경로 문제는 Bonet et al.[11]과 Yu et al.[18] 등의 연구가 있다. Bonet et al.은 에이전트가 비용을 최소화하면서 목적지에 도달하기 위한 경로를 도출하기 위해, 가치반복법(Value Iteration)과 실시간동적 계획법(Real-Time Dynamic Programming)으로 최적 정책(policy)을 도출했다. Yu et al.은 교통신호(traffic signal)로 인한 차량의 대기시간을 동적인 비용으로 보고, 목적지까지 이동 시간을 최소로 하는 각 구간에서의 최적 정책을 보여주었다.

마코프 의사결정 프로세스는 경로 문제 외에도 의사결정을 요하는 다양한 분야에서 사용되었다. Schaefer et al.[16]은 콩팥 및 심장 질환의 치료 시점을 결정하는데 마코프 의사결정 프로세스가 효과적인 방법이 될 수 있음을 보여주었다. 또한, Swarup et al.[17]은 전염병 통제를 위한 의사결정 방법론으로 게임이론과 더불어 마코프 의사결정 프로세스를 기반으로 한 방법론을 제안하였다. Alagoz et al.[10]은 간 이식수술을 시행함에 있어 어떤 단계에서 수술을 하는 것이 환자에게 가장 긍정적인 지를 Bellman 부등식을 통해 도출했다.

국내에서는 마코프 의사결정 프로세스 기반의 문제와 관련, 무인기의 임무 할당 및 정찰 경로 선택에 POMDP(Partially Observable MDP)가 적용된 연구가 있다[1]. POMDP는 기본적으로 MDP의 논리를 근간으로 하지만, 에이전트의 시야가 자신의 주변(관찰이 가능한 상태의 집합)으로 한정되어 있고 자신의 현 위치를 알지 못한다는 것을 전제로 하기 때문에 연산 과정이 MDP보다 다소 복잡하다. 해당 논문에서는 무인기가 가지는 보상에 대해 대공 미사일에 의한 격추(-), 목표물 정찰 성공(+), 임무를 완수하지 못할 경우 매시간당 보상 삭감(-) 등을 구조화함으로써 효과적인 임무 경로를 제시했다.

본 연구가 기존의 연구와 다른 점은, 첫째, 에이전트의 목적지를 하나의 상태에 국한하지 않고 다수를 대상으로 삼았다는 것이다. 즉, 에이전트의 궁극적 목적은 어떤 방향으로든 폐쇄된 공간을 벗어나는 것이다. 둘째, 다른 상태로 이동할 수 있는 방향에 대해 특정 셀이나 동 · 서 · 남 · 북 4방향이 아닌 6방향을 기본으로 했다. 이를 통해 현실 설명력을 높였을 뿐 아니라, 일정 시간이 지난 이후에 에이전트의 위치 파악을 용이하도록 했다. 셋째, 앞서 설명한 기존 확률론적 기반 연구에서는 최적 정책을 구해 목적지에 도달하는 경로만 보여주었으나, 본 연구에서는 최적 정책 뿐 아니라 일정시간 이후에 에이전트가 각각의 상태에 존재할 확률도 함께 제시했다. 특히, 확률을 도출하는 과정에서 매트릭스(matrix)의 특성을 이용해 복잡한 계산을 손쉽게 처리할 수 있는 방법도 함께 제시했다. 넷째, 지형정보체계(GIS; Geographic Information System)로

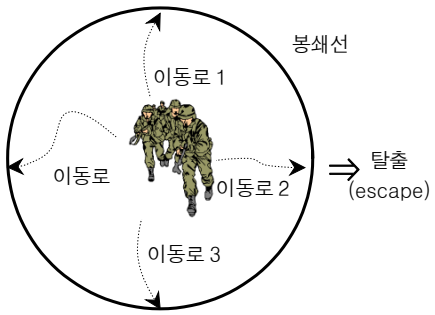
부터 추출한 정량적인 값으로 보상 구조(reward structure)를 구성해 객관성을 높였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 마코프 의사결정 프로세스를 적용해서 문제 정의 및 모델링을 하고, 제 3장에서는 사례연구 및 분석을 통해 모델을 현실에 적용함으로써 실효성을 검증한다. 마지막으로 제 4장에서는 본 연구의 의의와 한계점, 발전 방향에 대해 고찰해 본다.

2. MDP를 적용한 예측 모델

2.1 문제 정의

본 연구는 [그림 1]와 같이 폐쇄된 공간(이하 봉쇄구역)에 있는 에이전트가 그 공간에서 벗어나기 위해 임의의 방향으로 움직일 때, 에이전트의 이동 경로를 마코프 의사결정 프로세스를 이용해 예측하는 문제이다.



[그림 1] 봉쇄구역에서의 에이전트 이동 개념

에이전트 탐색자 입장에서는 에이전트를 잡는데 있어 최소의 인원과 시간을 투입하는 것이 바람직하다. 이를 위해서는 현 시점에서 에이전트가 존재할 확률이 높은 지역을 예측해야 하며, 일정시간 동안 에이전트를 잡지 못했다면 봉쇄구역을 더 확장시켜야 할지 여부를 결정해야한다. 요컨대, 본 연구는 에이전트 입장에서 가장 탈출하기 용이한 경로·지역이 어디인지 예측하는 모델이며, 연구에서 사용된 가정은 아래와 같다.

가정 1 : 탐색자로부터 추적을 당하고 있는 에이전트는 최초 봉쇄구역의 중앙에 위치하며, 봉쇄구역을 탈출하는 것을 목표로 한다. 이를 위해 일정한 속도로 쉬지 않고 움직인다.

가정 2 : 봉쇄선은 사람이나 장비가 지키고 있으므로 에이전트가 쉽게 탈출할 수 없다. 단, 탈출 가능성이 낮지만 탈출이 불가능한 것은 아니다.

가정 3 : 에이전트는 이동하는 과정에서 노출을 줄이고, 체력을 비축할 수 있으며, 평소에 경험하거나 훈련받은 지형을 선택해서 움직인다.

가정 4 : 에이전트는 이동 과정에서 자신이 선택한 지역으로 100% 이동할 수는 없다(체력 고갈, 방위 장비 고장, 탐색자 사전 선점 등이 원인).

2.2 MDP 기반 모델링

본 장에서는 제 2.1절에서 정의한 문제에 대해 MDP 기반 모델링을 한다. 연속적인 공간에서 에이전트의 이동이 가능한 공간은 무한대이므로 문제를 단순화하기 위해서 본 연구에서는 연구 대상 지역이 [그림 2]과 같이 육각형의 셀들로 이루어져 있으며 에이전트는 셀 중심에서 셀 중심으로 이동한다고 가정한다. 예를 들어, [그림 2]은 총 36개의 셀들로 구성된 지역을 나타내며, 굵은 실선은 봉쇄선을 의미한다.

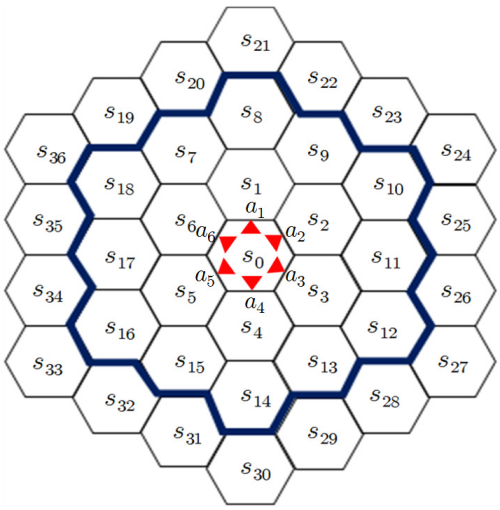
MDP 모델링을 위한 상태(state), 행동(action), 전이확률(transition probability), 보상(reward)은 다음과 같이 정의한다.

상태(state)는 봉쇄선을 기준으로 안쪽을 보통상태(normal state), 바깥쪽을 목적상태(object state)라고 정의하며, S_{normal}은 보통상태의 집합을, S_{object}는 목적상태의 집합을 나타낸다.

$$S = \{S_{normal}, S_{object}\}$$

예를 들어, [그림 2]에서 보통상태는 $s_0 \sim s_{18}$ 셀이

고, 목적상태는 s_{19}~s_{36} 셀이다. 에이전트는 봉쇄된 공간 중심에서 보통상태를 거쳐 이동, 궁극적으로 목적상태에 도달하는 것을 목표로 한다. 마코프 체인 관점에서 본다면 보통상태는 일시상태(transient state)를, 목적상태는 흡수상태(absorbing state)를 가리킨다고 보아도 무방하다.



[그림 2] 에이전트 이동 공간 분할

행동(action)은 각 셀에서 에이전트가 이동할 수 있는 6개 방향의 집합으로 이루어지며, A = {a_1, a_2, a_3, a_4, a_5, a_6}로 표현한다. [그림 2]에서 에이전트가 s_0 상태에 있다면, 6개의 정책을 선택할 수 있음을 알 수 있다. 단, 행동(선택)은 보통상태에서 이루어지는 것으로 만약, 에이전트가 봉쇄선을 통과해 목적상태에 도달했다면 최종 목표를 달성한 것이므로 더 이상 행동을 하지 않는다.

전이확률(transition probability)은 $P(s'_{normal}|s_{normal})$와 $P(s_{object}|s_{normal})$ 2종류로 구분된다. (s_{normal}, $s'_{normal} \in S_{normal}$, $s_{object} \in S_{object}$, $s_{normal} \neq s'_{normal}$) $P(s'_{normal}|s_{normal})$은 최적 정책으로 선택된 방향으로 이동할 확률을 75%, 나머지 5방향으로 이동할 확률을 5%라고 가정한다. 에이전트가 최적 정책으로 선택한 지역으로 이동할 확률을 정하기 위해, 이와 유사한 군사훈련에서 데이터를 참고하였다. ○○○ 부대 데이터에 따르면 방향 유지가 우수한 침투부대가 목표 지역으로 정확히 이동할 확률은 약 90%였으며, 상대적으로 훈련이 미흡한 침투부대의 경우, 약 73%의 방향 유지율을 보였다. 본 연구에서 에이전트는 고도로 훈련받은 요원이지만 탐색자에게 포위된 불리한 상황임을 가정, 가장 미흡한 부대의 방향 유지율 수준을 적용했다. 또한, 에이전트가 원하는 방향으로 이동하지 못했을 경우, 나머지 방향으로의 이동 확률을 동일하게 적용해 불확실성을 높임으로써 탐색자 입장에서 가장 보수적인 상황을 가정했다. [그림 2]에서 s_0에 있는 에이전트가 최적 정책으로 a_1을 선택했다면, 다음 스텝에서 에이전트가 s_1에 있을 확률이 75%, s_2~s_6에 있을 확률이 각각 5%가 된다.

$P(s'_{object}|s_{normal})$은 1.5%로 가정했다. 이는 봉쇄구역 내에 있는 에이전트가 봉쇄선을 통과할 확률이므로 $P(s'_{normal}|s_{normal})$에 비해 상대적으로 낮다고 보는 것이 합리적이다. 하지만, 탐색자가 봉쇄선을 어떤 방식으로 구축하느냐에 따라 달라질 수 있기 때문에 의사결정자가 선택할 문제라고 할 수 있다. 즉, 병력이나 장비를 얼마나, 어떻게 배치하느냐에 따라 봉쇄선을 통과할 확률은 더 높아질 수도, 낮아질 수도 있는 것이다.

보상(reward)은 각 상태에서 에이전트의 결정에 영향을 미치는 유인이다. 따라서, 그 종류에 따라 수준이 상이하다. 에이전트의 목표는 목적상태로 이동하는 것이므로 중간 과정의 성격을 가진 보통상태보다 보상값이 상대적으로 더 크다. 본 연구에서는 지형정보체계 전자지도로부터 보상구조와 관련된 정보를 추출해 객관성을 높였으며, 이 과정에서 상용 툴인 ArcGIS 10.1 프로그램을 사용했다.

우선, 보통상태에서의 보상은 은닉 요인, 체력 요인, 전술 요인을 기준으로 결정된다고 가정하며, 각 요인은 에이전트의 선택에 있어 복합적으로 작용하는 정량적인 값이다.

은닉 요인은 에이전트가 숨어 다니기 용이한 정

도를 나타내며, 인적이 드물고 은신하기 쉬운 곳일수록 값이 크다. 특정 지역(상태)에서 인적이 많은지 적은지는 도로, 실측건물, 경작지, 분기점, 우마차로(차량이 다닐 수 있는 정도 산길)의 존재 여부에 따라 결정되며, 가중치는 <표 1>과 같다[5]. 은닉 요인에 영향을 미치는 또 다른 요소는 식생(산림의 우거짐 정도)과 암석지대 여부이다. 즉, 식생의 밀집도가 높고 암석지대가 아니라면 보상값이 커진다. 식생은 GIS 체계의 전자지도상에서 88, 79, 72, 65, …, 0의 이산적인 값으로 표시되는데, 88은 산림이 우거진 정도가 가장 높고 0은 해당 지역에 산림이 존재하지 않음을 의미한다.

〈표 1〉 인적 결정 요소 가중치

도 로	건 물	경작지	분기점	우마차로
0.3	0.3	0.2	0.1	0.1

체력 요인은 에이전트가 체력을 유지하는데 있어 유리한 지역인지 여부를 가늠하는 척도이다. 도망자 입장에서 체력이 고갈되어 더 이상 은닉하거나 봉쇄선 돌파 시도를 할 수 없게 되므로 에이전트가 이동을 하면서도 체력을 얼마나 유지, 비축할 수 있는가는 다음 이동지역을 선택하는 중요한 기준이 된다. 체력 요인은 경사가 완만하고 짐승들이 이동하는 길(짐승로)이 있거나 급류가 없는 지역일수록 긍정적인 영향을 미친다[5].

〈표 2〉 능선별 가중치

6부	5·7부	4·8부	3·9부	2·10부	1부
1	0.8	0.6	0.4	0.2	0

전술 요인은 에이전트가 평소 훈련을 받거나 익숙한 지형인지 여부를 나타낸다. 일반적으로 산악에서 도주하는 에이전트(무장공비)는 5~7부 능선을 따라 이동하도록 훈련을 받는다. 따라서, 각 능선(고도)별로 상이한 보상값을 가지는데, 능선별 가중치는 <표 2>와 같다[4].

최종적으로 도출되는 보상값은 앞서 설명한 3가지 요인값의 곱으로 나타내며, 0~1사이의 값을 가진다. <표 3>은 임의의 보통상태에서 보상값을 도출하는 예를 보여준다. 은닉요인 보상값은 1-{0.3(도로)+0.2(경작지)+분기점(0.1)}×식생(0.7) = 0.35, 체력요인 보상값은 0.8(경사)×0.7(소로)×0.7(급류) = 0.392, 전술요인 보상값은 1(고도, 6부 능선)이며, 최종적인 보상값은 0.35(은닉보상)×0.392(체력보상)×1(전술보상) = 0.1372로 계산된다.

목적상태의 보상값은 본 MDP 모델에 있어 에이전트가 봉쇄구역을 돌파하고자 하는 의지를 산술적으로 나타낸다. 목적상태의 보상값이 보통상태의 보상값과 비슷한 수준이라면, 에이전트는 굳이 목적상태로 이동하려하지 않는다. 한편, 목적상태의 보상값 설정 방식에 대해서는 기존에 연구된 바가 없으므로 본 연구에서는 목적상태의 보상값을 10, 20, 30, …으로 증가시키면서 실험을 진행했다.

감가인수(discount factor)는 가치반복법으로 각 상태의 효용(utility)과 최적 정책을 도출할 때 그 값이 수렴하도록 해 주는 기능[12]을 하며, 본 연구에서는 0.9를 적용했다.

본 모델에서 구현하고자 하는 목표는 2가지로, 각 상태에서 에이전트가 어떤 방향으로 이동할지 선택하는데 필요한 최적 정책(π)과 특정 시점 이후에 각 상태에서 에이전트가 존재할 확률을 도출하는 것이다. 최적 정책은 위에서 제시한 MDP 요소를 바탕으로 Bellman 부등식을 이용해 가치반복법으로 도출했다. 한편, 각 상태에서 에이전트가 존재할 확률은 아래와 같이 도출했는데, 최적 정책이 정해지면 이를 매트

〈표 3〉 상태 정보를 바탕으로 보상값 도출(예)

은 닉 요 인								체 력 요 인				전 술 요 인		보상값
도로	건물	경작지	분기점	우마차로	암석지대	식생	은닉보상	경사	소로	급류	체력보상	고도	전술보상	
○	×	○	○	×	×	79	0.35	4.65	×	○	0.392	816.9	1	0.1372

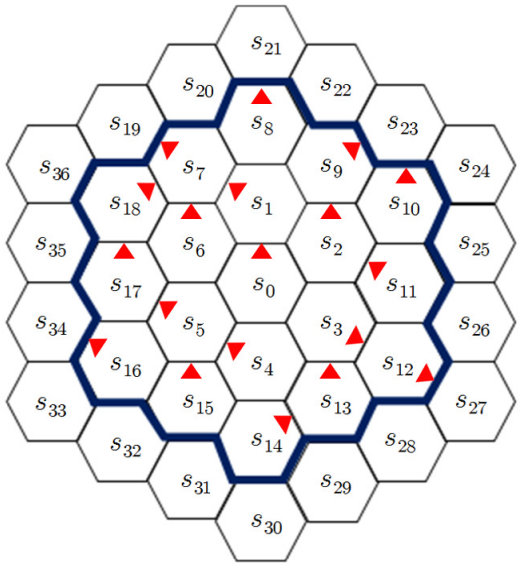
릭스의 특성을 활용해 구하는 방법을 아래와 같이 제시한다.

예를 들어, 상태가 37개인 축소된 모델에서 최적 정책이 [그림 3]의 삼각형과 같고, 0스텝(t = 0)에서 에이전트는 s_0에 있다고 가정한다.

0스텝의 최적 정책은 $s_0 \rightarrow s_1$ 방향이므로, 1스텝 이후(t = 1)에 에이전트는 s_1에 75%, s_2, s_3, s_4, s_5, s_6에 각각 5%의 확률로 존재할 것이므로, 아래와 같은 매트릭스로 표현할 수 있다.

1st 스텝 : $\wp_1$ (1×37 matrix) =

	0	1	2	3	4	5	6	7	···	36
0		0.75	0.5	0.5	0.5	0.5	0.5			



[그림 3] 각 상태에서의 최적 정책

여기서 $\wp_1$은 1 by 37 매트릭스인데, row number인 0은 최초 에이전트가 s_0에 있음을 의미하고, column number 1, 2, 3, 4, 5, 6은 각각 s_0의 인접 상태를 나타낸다. 즉, 0스텝에서 1스텝으로 전이된 이후에는, $\wp_1$의 column number에 에이전트가 전이될 확률이 들어가 있는 것을 볼 수 있다.

여기까지는 직관적으로 알 수 있을 정도로 간단히 계산할 수 있지만, 2스텝(t = 2)만 되더라도 계산이 기하급수적으로 복잡해짐을 알 수 있다. 한편, 1스텝 이후에 에이전트가 가지는 최적 정책을 매트릭스로 표현하면 아래와 같다.

Policy at 1st step $\wp_{policy1}$(37×37 matrix) =

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	···	36
0		0.75	0.05	0.05	0.05	0.05	0.05														
1	0.05		0.05				0.05	0.75	0.05	0.05											
2	0.05	0.05		0.05						0.75	0.05	0.05									
3	0.05		0.05		0.05							0.05	0.75	0.05							
4	0.05			0.05		0.75								0.05	0.05	0.05					
5	0.05				0.05		0.05									0.05	0.05	0.75			
6	0.05	0.05				0.05		0.75										0.05	0.05		
⋮																					
36																					

$\wp_{policy1}$은 1스텝→2스텝으로 전이될 때 MDP 모델에 의해 도출된 최적 정책에 따라 에이전트가 각 상태로 이동할 확률을 나타낸 매트릭스이다. 따라서, $\wp_1$과 $\wp_{policy1}$의 곱을 통해 $\wp_2$를 도출할 수 있다.

$$2^{nd} \text{ 스텝 } : \wp_2 = \wp_1 \times \wp_{policy1}$$

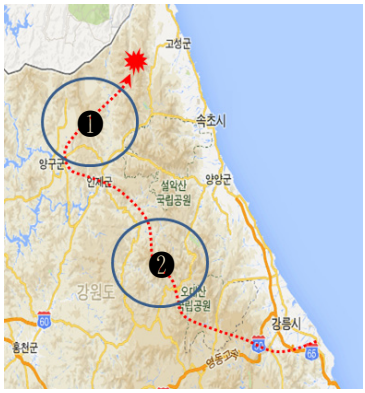
$\wp_2$ 역시 1 by 37 매트릭스이며, 마찬가지로 column number를 보면 어떤 상태에서 에이전트가 존재할 확률이 얼마인지 알 수 있다. 이 같은 방법으로 아래와 같이 n스텝 이후에 각 상태에서 에이전트가 존재할 확률을 도출할 수 있다.

$$3^{rd} \text{ 스텝 } : \wp_3 = \wp_2 \times \wp_{policy2}$$
$$4^{th} \text{ 스텝 } : \wp_4 = \wp_3 \times \wp_{policy3}$$
$$\vdots$$
$$n^{th} \text{ 스텝 } : \wp_n = \wp_{n-1} \times \wp_{policy\ n-1}$$

3. 사례연구 및 분석

모델을 구현하기 위해, Intel(R) Core(TM) i53470 CPU 3.20GHz, 16GB PC를 사용했으며, 프로그램은 java eclipse, luna 버전을 사용했다. 본 모델은 임의

의 도망자를 대상으로 한 일반적인 모델인 바, 그 유효성을 검증하기 위해서는 실증 적용이 반드시 필요하다. 여기서 유효성이란, 현실에서 동일한 상황에 놓여있는 에이전트가 취하는 행동을 모델이 얼마나 잘 예측할 수 있는가를 의미한다. 이에 1996년 강릉 무장공비침투사건 당시 실제 무장공비들이 이동한 경로 2곳을 바탕으로 모델의 예측력을 검증하였으며, 그 위치는 [그림 4]과 같다.



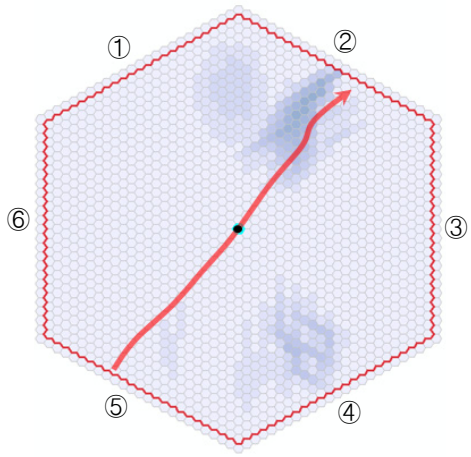
[그림 4] 모델 적용 대상 경로

본 사례에 모델을 적용하기 위해 ❶과 ❷지역 상태의 개수를 각각 1,801개로 분할했다. 보통상태 1,657개, 목적상태 144개로 구성되며, 전체 봉쇄구역의 넓이는 $478{,}334.475m^2$이다.

우선, ❶지역에 해당하는 지역에 해당하는 지역을 모델에 적용한 결과는 [그림 5]과 같다. 봉쇄선 내부에서 좌하에서 우상으로 이어진 선은 실제 무장공비가 이동한 경로를 나타낸다.

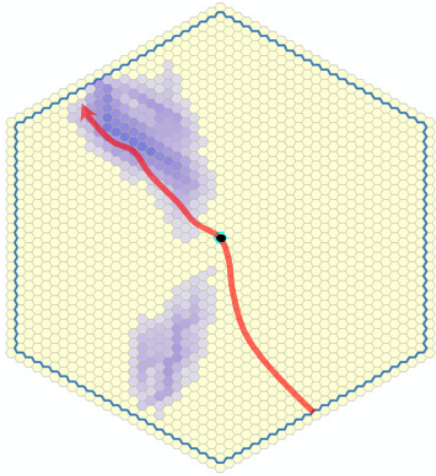
단, 북쪽으로 이동하려는 무장공비의 심리를 반영하기 위해, 동(③)·서(⑥)·남쪽(④⑤)에 비해 북쪽(①②)에 위치한 목적상태의 보상을 좀 더 크게 (+10) 부여했다.

[그림 5]에서는 에이전트가 봉쇄구역 중앙에서 봉쇄선까지 도달하는 최단시간인 23스텝에서 에이전트가 존재할 확률이 높은 상태(지역)을 표시하였다. 즉, 색깔이 진할수록 에이전트가 존재할 확률이 높은 상태(지역)임을 나타낸다.



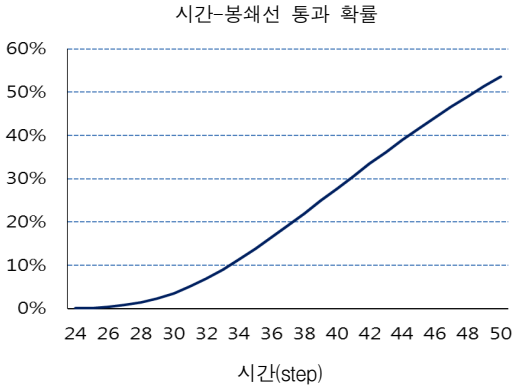
[그림 5] ❶지역 : 23스텝 후 에이전트 존재 확률

❷지역에 해당하는 지역에 해당하는 지역을 모델에 적용한 결과는 [그림 6]과 같다. 마찬가지로 23스텝 이후 에이전트가 존재할 확률이 높은 상태를 음영으로 표시했다. 결과에서 알 수 있듯이 본 모델을 실제 무장공비의 이동 경로에 적용해 본 결과, 특정 시간 이후에 에이전트가 존재할 확률이 높은 지역과 무장공비의 실제 이동 경로가 거의 일치하는 것을 볼 수 있다. ❷지역을 기준으로 시간의 흐름에 따라 에이전트가 봉쇄선을 통과, 봉쇄구역을 벗어났을 확률은 [그림 7]와 같다.



[그림 6] ❷지역 : 23스텝 후 에이전트 존재 확률

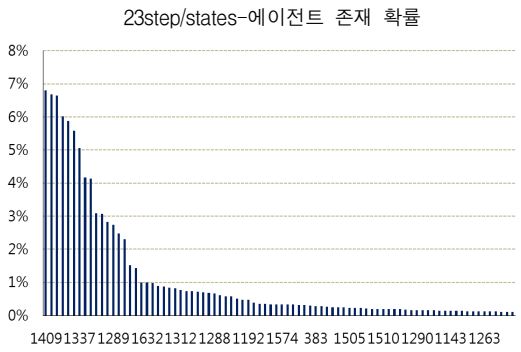
에이전트가 봉쇄선을 통과할 확률은 24스텝부터 나타나기 시작해 48스텝에 이르러서는 50% 가까이 증가한다. 이 같은 비율로 100스텝까지 이르게 되면 에이전트가 봉쇄선을 통과했을 확률은 96.215%가 된다. 즉, 100스텝이 지날 때까지 에이전트를 잡지 못했다면 봉쇄구역 내에서 아무리 찾아봐야 잡을 확률이 거의 없음을 알 수 있다.



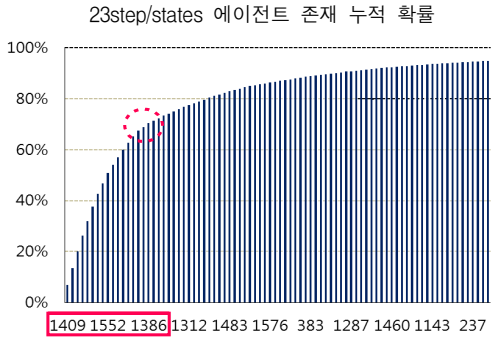
[그림 7] 에이전트의 봉쇄구역 탈출 확률

따라서, 봉쇄구역을 넓히는 것이 합리적인 수색 방법이 될 것이다. 만약 시간에 따른 봉쇄선 통과 확률을 더 줄이려면 봉쇄선을 지금보다 더 강력하게 구축해야 한다. 즉, 더 많은 인원과 장비를 투입해 봉쇄선을 구축한다면 각 상태에서 에이전트가 봉쇄선을 통과할 확률은 지금보다 낮아질 것이며, 시간의 흐름에 따른 에이전트의 봉쇄구역 탈출 확률도 그 만큼 줄어들 것이다. 다만 이 경우에는 더 많은 비용이 소요된다. 실제 무장공비 수색 작전 시에도 포위망을 강하게 구축하는 것이 중요하지만 그 범위를 넓힐수록 인원과 비용이 기하급수적으로 증가한다.

의사결정자는 본 모델을 통해 봉쇄구역 유지, 확대 여부 뿐 아니라 탐색자 투입 지역과 관련해서도 효과적인 정보를 얻을 수 있다. 즉, 처음으로 무장공비가 목격(신고)된 시점으로부터 지금까지 얼마만큼의 시간이 흘렀다면, 현 시점에서 부장공비를 잡기 위해 어느 곳에 병력을 투입해야 할 것인지 쉽게 판단할 수 있다. 23스텝 이후에 무장공비가 존재할 확률이 높은 상태(지역)를 내림차순으로 표시하고 그 확률이 높은 지역부터 수색하는 것이 합리적인 방법이라 할 수 있다. ❷지역, 23스텝에서 무장공비가 존재할 확률이 높은 상태를 내림차순으로 정리하면 [그림 8]과 같다. 여기서 x축은 상태의 인덱스(index)를, y축은 해당 상태에서 무장공비가 존재할 확률을 의미한다.



[그림 8] 에이전트의 존재 확률이 높은 상태



[그림 9] 에이전트의 존재 확률이 높은 상태

하지만, 이 그래프만으로는 효과적인 무장공비 병력 투입 지역을 파악하는 데는 애로가 있다. 물론, 확률이 높은 상태(지역)부터 병력을 투입하는 것이 합리적이라 하더라도, 과연 그 효과가 어느 정도 인지는 알기 어렵다. 이에, [그림 8]의 확률을 누적시킨 그래프 [그림 9]을 제시한다. 합리적인 군 지휘관이라면 무장공비가 있을 것으로 예상되는 상태의 누적확률이 70% 정도가 되는 지역에 병력을 우선 투입할 것이다. 제한된 병력을 이 지역에 집중할 경우,

그 만큼 무장공비를 잡을 확률이 커지기 때문이다. 반면, 나머지 30%의 확률에 해당하는 지역을 수색하기 위해서는 상대적으로 병력을 광범위한 지역에 투입해야하므로 효율성이 급격히 떨어진다고 예상할 수 있다.

즉, 전체 수색 지역의 1%에 해당하는 부분(1,801개 상태 중 16개 상태)에 병력을 투입, 수색한다면 무장공비를 잡을 확률이 70% 정도가 될 것이다. 만약, 특정 상태에 투입된 병력이 무장공비를 탐지할 확률이 α라면, 실 탐지 확률은 $0.7\times\alpha$이다(α는 수색 병력, 장비, 수색자의 숙련도 등에 따라 좌우) 물론, 수색 병력이 충분해 모든 지역에 병력을 투입할 수만 있다면 좋겠지만, 현실에서 사실상 불가능한 일이다.

결과적으로 본 모델은 실제 무장공비의 이동경로를 통해 그 예측력과 효과를 입증하였다. 에이전트가 합리적인 의사결정자라면, 이 같은 결과가 나타나는 것은 어찌 보면 당연하다. 도망자가 봉쇄선을 벗어나기 위해서 상황을 조망해 각 상태에서 최선의 행동을 취할 것은 자명한데, 본 모델은 이 같은 현실의 모습을 충실히 구현했기 때문이다.

4. 결 론

본 논문에서는 도망자인 에이전트가 봉쇄구역에서 벗어나기 위해 의사결정을 내리는 과정을 MDP를 이용해 모델링했다. 특히, GIS로부터 추출한 데이터를 이용해 보상 구조를 구성함으로써 객관성을 높였다. 특히, 기존 모델에 비해 이동 방향이 다양하고, 한 지점에서 출발해 여러 개의 목적지 중 어느 곳으로든 도착할 수 있는 새로운 형태의 모델이라는 점에서 의의가 있다. 또한, 매 스텝별 각 상태에서 에이전트가 존재할 확률을 도출함에 있어, 매트릭스의 특성을 이용한 방법을 새롭게 제시했다. 또한 모델을 현실에 적용해 본 결과, 상당히 높은 예측력을 가지는 것으로 확인되었다. 모델의 유효성을 검증하기 위해 1996년 강릉 무장공비침투사건 당시 이동 경로를 대상으로 했지만, 비단 이같은 문제 뿐 아니라 범죄자(탈영병) 추적, 북한 잠수정 및 항공기 침투경로 예측, 포위된 적군의 이동로 예측 등 군사 분야는 물론, 교통신호 알고리즘, 물류 이동 등 다양한 분야에 광범위하게 활용될 수 있을 것으로 기대된다.

현재 검토하고 있는 후속 연구는 다음과 같다. 첫째, 에이전트가 최적 정책을 선택했을 때의 이동 확률을 동적으로 부여하는 문제이다. 본 연구에서는 시간의 흐름에 관계없이 최적 정책 선택 방향으로 75%의 확률로 이동한다고 가정했다. 하지만 이 확률은 시간에 따라 변할 가능성도 존재한다. 즉, 에이전트를 제약하는 요인이 상대적으로 크지 않은 초기에는 에이전트가 자신이 선택한 방향으로 쉽게 이동하겠지만, 시간이 지남에 따라 점차 그 확률은 감소될 소지가 있다. 도망자인 에이전트가 조기에 봉쇄구역을 벗어나지 못했을 경우, 식량·체력 고갈, 탐색자의 증가, GPS 장비의 고장 등으로 인해 행위에 제약이 가해질 가능성이 높기 때문이다. 둘째, 보상이 동적으로 부여되는 상황에 대한 연구이다. 에이전트가 은닉해서 이동하던 중 어떤 상태에 탐색자가 있는 지를 관찰(확인)했다면, 그 상황에서 다시 가치반복법을 시행해서 각 상태의 효용을 재산출 할 것이다. 그 이후에는 최적 정책이 재수정될 소지가 있다. 만약, 탐색자 존재 여부에 따라 보상이 변화될 수 있도록 모델링한다면, 좀 더 복합적이고 다양한 상황을 묘사할 수 있을 것이다.

참 고 문 헌

[1] 김동호, 이재송, 최재득, 김기응, “복수 무인기를 위한 POMDP 기반 동적 임무 할당 및 정찰 임무 최적화 기법”, 「정보과학회지」, 제39권, 제6호(2011), pp.453-463.

[2] 민대기, “추계 계획법을 이용한 수술실 약 모델과 Newsvendor 비율의 자원 효율성에 한 향 분석”, 「경영과학」, 제28권, 제2호(2011), pp.17-29.

[3] 방수남, 허 준, 손홍규, 이용웅, “지형공간정보 및 최적탐색기법을 이용한 최적침투경로 분석”, 「대

한토목학회논문집」, 제26권, 제1D호(2006), pp. 195-202.

[4] 신내호, 오명호, 최호림, 정동윤, 이용웅, “지형 공간정보 기반의 침투위험도 예측 모델을 이용한 최적침투지역 분석”, 「한국군사과학기술학회지」, 제12권, 제2호(2009), pp.199-205.

[5] 육군본부, “대침투작전 전투기술”, 「야전교범 3-0-2」, 2009.

[6] 윤봉규, “전장 모델링 실무자를 위한 마코프 체인에 대한 소고”, 「국방과학기술」, 제2권, 제3호(2009), pp.47-61.

[7] 이건창, 한민희, 서영욱, “탐색 및 활용을 통한 컴퓨터 매개 커뮤니케이션의 팀 창의성에 관한 연구 : 에이전트 모델링 기법을 중심으로”, 「경영과학」, 제28권, 제1호(2011), pp.91-105.

[8] 정석윤, 허 선, “마코프 재생과정을 이용한 ATM 트랙픽 모델링 및 성능분석”, 「한국경영과학학회지」, 제24권, 제3호(1999), pp.83-91.

[9] Abhijit, G., “One Step Sizes, Stochastic Shortest Paths, and Survival Probabilities in Reinforcement Learning,” *Proceeding of the 40th Conference on Winter Simulation. Winter Simulation Conference*, 2008.

[10] Alagoz, O., H. Hsu, A.J. Schaefer, and M.S. Roberts, “Markov Decision Processes : A Tool for Sequential Decision Making under Uncertainty,” *Medical Decision Making*, Vol. 30, No.4(2010), pp.474-483.

[11] Bonet, B. and H. Geffner, “Solving Stochastic Shortest-Path Problem with RTDP,” *Technical report, University of California, Losangeles*, 2002.

[12] Hyeong, S.C., C. M. Fu, H.J. Hu, and M.I. Steven, “Simulation-Based Algorithms for Markov Decision Processes,” *Springer,* 2006.

[13] Kolobov, A., Mausam, and D.S. Weld, “Stochastic Shortest Path MDPs with Dead Ends,” *HSDIP*, Vol.78, No.10(2012), pp.78-86.

[14] Pan, Y., L. Sun, and M. Ge, “Finding Reliable Shortest Path in Stochastic Time-Dependent Network,” *Procedia-Social and Behavioral Sciences*, Vol.96, No.6(2013), pp.451-460.

[15] Ravindra, K.A. and L.T. Magnanti, B.J. Orlin, “Network Flows,” *Prentice Hall*, 1993.

[16] Schaefer, A.J., M.D. Bailey, S.M. Schechter, and M.S. Roberts, “Modeling Medical Treatment using Markov Decision Processes,” *Handbook of Operations Research/Management Science Applications in Health Care*, *Kluwer Academic Publisher*, 2004.

[17] Swarup, S., G. Eubank, and M.V. Marathe, “Computational Epidemiology as Challenge Domain for Multiagent Systems,” *Proceeding of International Conference on Autonomous Agents and Multi-agent Systems. International Foundation for Autonomous Agents and Multi-agent Systems*, 2014.

[18] Yu, X.-H. and W.W. Recker “Stochastic Adaptive Control Model for Traffic Signal Systems,” *Transportation Research Part C*, Vol.14, No.4(2006), pp.263-282.