Sequential Decision 理論과 應用

陸軍少領 黃 東 準*

I. 序 論

Sequential decision 理論은 在庫統制, 待期線(waiting lines), 維持(maintenance) 및 代替(replacement), 信賴(reliability), marketing, 財務管理, 그리고 消費者 行態分析 問題등의 經營經濟體系(business and economic system)分析에 대단히 많이 應用되고 있다. 體系(system)의 動的인 關係를 分析함에 있어서 最適化(optimization)와 體系 行態(behavior)의 關係를 綜合하여 全體的인 計劃 및 統制를 할 수 있는 解決案(solution)을 주는 技法이 바로 Sequential decision 理論이며 通常 Markov decision process (MDP)라고 稱하기도 한다. Markov decision prorcss 는 i) 體系의 狀態空間(state space)S, ii) 行動空間(action space)A, iii) 移動法則(law of motion)T, 그리고 iv) 費用收益構造(reward structure)W에 의하여 特定지워진다. 여기에 變移時間(transition time)을 考慮하는 경우 Semi-markov decision process라고 한다.

一般的으로 體系의 行態變化는 時間的으로 連續的이나 大部分의 經營經濟模型들은 주어진 不連續時間(discrete time)으로 連續時間過程(continuous time process)에 대한 推定을 可能하게 하며, 이렇게 하여 얻어진 結果는 質的으로나 量的으로 連續時間過程에 의하여 얻은 結果와 같다. 體系의 狀態가 觀察되고 그 狀態에 대한 어떤 動作이나 決定이 취해졌을 때 體系의 狀態는 變하게되고 이러한 過程은 循還된다. 계속적인 decision process가 體系의 狀態를 變하게할 때 limit가 存在한다면 體系의 行態를 좀 세밀하게 分析할 수 있을 것이다. 그러나 連續時間過程에서의 limit 存在여부는 sequential decision 理論에서 重要한 問題로 模型에 대한 解決案의 存在(existence)여부 등 分析에 어려움이 많다. 이러한 limit 存在에 대한 어려움을 없애기 위해 흔히 有限狀態(finite state), 有限의 行動數(finite number of action), 그리고 不連續時間(discrete time)으로 體系를 模型化한다.

이렇게 discrete time process에 의한 模型이 continuous time process의 推定解라는 事實外에 우리주변의 經營經濟模型들은 discrete time process가 大部分이고, 더구나 分析에 必要한 計算은 digital computer에 의하여 쉽게 處理할 수 있다는 理由때문에 finite state markov chain은 continuous state continuous time process 보다는 더 많이 應用되고있다.

여기에서는 一般的인 sequential decision 理論의 基本的 槪念과 Howard와 Bellman의 問題解決方法(algorithms)들을 提示하고, marketing 問題中에서 여러 廣告代案에 대한 效果分析과 거기에 따라 最適의 政策方案을 決定하는 過程을 sequential decision 理論으로 보이고자 한다.

II. Sequential decision 理論

1. 基本動態(basic dynamics)

Sequential decision(markov decision process)은 Markov chain의 條件的確率(conditional probability)의 一般化를 必要로 한다. 便宜上 體系가 가질 수 있는 狀態의 數를 m개라 한다. 즉 狀態空間 $S=\{1, 2, 3\cdots, m\}$ $T_t(j|i,a)$를

* 陸軍士官學校

時間$(t-1)$에 體系가 狀態 i에 있고 行動a를 취했을 때 體系가 時間 t에 狀態 i에 있을 確率이라 한다. process가 m개의 狀態$(1, 2, \cdots, m)$을 가지고 있으므로 $T_t(1|i,a)$, $T_t(2|i,a)$, $\cdots$, $T_t(m|i,a)$는 m차원의 row vector를 形成한다. 즉

$$T_t(i,a)=[T_t(1|i,a),\ T_t(2|i,a),\ \cdots,\ T_t(m|i,a)]$$

다음에는 各狀態에서 취하는 決定(action)에 대하여 살펴보자. $d_t(i)$를 時間$(t-1)$에 各 狀態 $i\in S$에서 취하는 行動을 나타내는 決定(decision)이라 하자. 다시 말하면 함수 d_t는 狀態空間으로부터 行動空間으로 옮겨주는 mapping function이다. d: $S\rightarrow A$. 變移確率(transition probability) vector들 $T_t(1,d_t(1))$, $T_t(2,d_t(2))$, $\cdots$, $T_t(m,d_t(m))$은 變移行列(transition matrix) $T_t(d_t)$를 形成한다. 이 變移行列 $T_t(d_t)$는 決定 d_t와 함께 體系의 移動法則을 나타낸다.

P_t를 時間 t에 體系가 여러 狀態에 있을 確率들의 vector라 하자. 그러면 體系의 動的 行態의 變化는 다음式으로 表示된다.

$$P_t=P_{t-1}T_t(d_t) \tag{1}$$

政策(policy)D를 各 期間에 취해진 一聯의 決定들 $d_1, d_2, \cdots, d_t$의 集合이라 하자. P_t는 政策D의 함수이고 이러한 一聯의 決定에 대한 依存性은 時間 o부터 t까지의 狀態變化에 대하여 變移行列들을 計算할 때 꼭 考慮되어야 한다. 이것은 反復的으로(recursively) 다음과 같이 計算된다.

$$T^{(t)}(D)=T^{(t-1)}(D)T_t(d_t)$$
$$T^{(o)}(D)=I \tag{2}$$

또는

$$T^{(t)}(D)=\prod_{n=1}^{t}T_n(d_n) \tag{3}$$

(2)와 (3)의 行列式들로 Dynamic law (1)式은 다음과 같이 다시 표시할 수 있겠다.

$$P_t(D)=P_0T^{(t)}(D) \tag{4}$$

(4)式의 경우는 一定한 變移行列을 가진 경우보다는 훨씬 複雜하다. 만약 變移行列T가 時間에 關係없이 一定하다면 $T^{(t)}$는 t回 累乘(t^{th} power)이다. One step 變移確率과 K번째 決定 d_k는 모두 時間에 依存하기 때문에 $T^{(t)}(D)$는 時間 o부터 時間 t까지의 可能한 體系의 狀態變化만을 나타낸 것이다.

體系의 狀態가 變함에 따라 費用과 收益이 發生하게 된다. $W_t(j,i,a)$를 時間$(t-1)$에 體系의 狀態 i에서 行動 $a\in A$를 취했을 때 時間 t에 狀態 j로 될 때 얻어지는 利益이라 하자. 體系 狀態가 i이고 行動 $a\in A$가 취해졌을 때 時間$(t-1)$과 t 사이에 얻어지는 豫想利益(expected gain) $W_t(i,a)$는 다음과 같다.

$$W_t(i,a)=\sum_{j=1}^{m}W_t(j,i,a)\cdot T_t(j,i,a) \tag{5}$$

여러狀態 $i\in S$에서 時間$(t-1)$에 使用된 決定 d_t에 의하여 行動들이 나타내졌을 때 期間 t에 얻어지는 豫想利益들의 vector $W(d_t)$는 다음과같이 表示된다.

$$W(d_t)=[Wt(1,\ d_t(1)),\ Wt(2,\ d_t(2)),\ \cdots,$$
$$Wt(m,\ d_t(m))]$$

따라서 政策 D를 使用하여 期間 t에 얻어지는 豫想利益은 아래式의 inner product이다.

$$(P_{t-1}(D),\ W_t(d_t)) \tag{6}$$

(6)式으로부터 體系를 運用하는 데 보통 다음과같은 세가지 尺度를 使用한다.

첫째 政策 D의 價値를 (6)式의 豫想收益들의 割引된 合計(discounted sum)로 計算하는 것이다. 즉 割引係數(discount factor) $\beta(0\leq\beta<1)$를 使用하여 割引된 合計는

$$\sum_{t=1}^{\infty}\beta^{t-1}(P_{t-1}(D),\ Wt(d_t)) \tag{7}$$

로 나타내진다. 모든 $W_t(j,i,a)$가 有限의 값을 가질 때 함수(7)은 반드시 存在한다.

둘째는 (6)式의 期間當 平均値를 구하는 것이다. 期間當 平均値는

$$\frac{1}{N}\sum_{t=1}^{N}(P_{t-1}(D),\ W_t(d_t)) \tag{8}$$

로 표시되나 割引된 合計의 경우처럼 함수(8)의 存在여부는 상당한 分析을 要한다. 다시 말하면 만약 N이 무한대로 가면서 limit가 使用되었다면 함수 (8)의 存在는 매우 複雜한 것이다.

세번째는 stationary 政策을 假定하는 경우다. stationary 政策은 各 期間에 항상 同一한 決定을 취하는 것을 말하며 決定함수 d에 대하여 stationary 政策은 $D=d^{\infty}=d,d,d,\cdots$로 表

示된다. 또한 時間에 따라 $W(d_t)$가 一定하다고 본다면,

$$\lim_{t\to\infty}(P_{t-1}(D),\ W_t(d)) \tag{9}$$

은 存在할 수도 있다. 둘째나 세번째의 경우와는 달리 첫번째의 割引된 合計에 의한 尺度는 함수의 存在여부가 問題가 안되기때문에 흔히 分析의 觀點에서 많이 使用된다.

2. Algorithms

앞에서 보았던 基本動態關係들은 政策의 價値를 評價하는 데 反復的으로 使用된다. 물론 代案의 政策들을 비교 分析하는 方法에는 여러가지가 있겠으나 더 效果的이고 體系的인 方法은 最適化를 反復關係(recursive relation)에 통합하는 것이며 이것이 여기에서 提示하는 Bellman과 Howard의 技法들이다.

1) Bellman의 Value iteration method.

이 方法은 보통 Successive Approximation Method라고도 하며 有限의 期間에 대한 最適의 政策을 發見하려고할 때 使用된다.

(4)式과 (6)式으로부터 t번째 期間에 있어서 예상수익에 基本관계의 다음 두가지 表現은 같은 것이다.

$$(P_oT^{(t-1)}(D),\ W_t(d_t)) = (P_o,\ T^{(t-1)}(D)\,W_t(d_t)) \tag{10}$$

(10)式의 反復關係는 K期間까지의 豫想收益

$$\sum_{t=1}^{K}(P_oT^{(t-1)}(D),\ Wt(d_t)) = (P_o,\ \sum_{t=1}^{K}T^{(t-1)}(D)\cdot W_t(d_t)) \tag{11}$$

를 계산하는 데 應用되며 아래와 같이 反復的으로 계산함으로써 K기간의 예상수익은 구해진다.

$$\begin{aligned} F_0&=0\\ F_{n+1}&=W_{k-n}(d_{k-n})+T_{k-n}(d_{k-n})F_n \end{aligned} \tag{12}$$

(12)를 계산하고 (P_0, F_k)를 계산함으로서 K예상수익은 아주 쉽게 얻어진다. 만약 割引된 合計를 計算하자면 $(0\leq\beta<1)$,

$$F_{n+1}=W_{k-n}(d_{k-n})+\beta T_{k-n}(d_{k-n})F_n$$

을 계산하면 된다.

最適의 政策 D를 구하기 위하여서는 F_{n+1}값을 최대화하는 各 期間에서의 決定들을 구해야 한다. 즉

$$\begin{aligned} H_0&=0\\ H_{n+1}&=\underset{d_{k-n}}{\mathrm{Max}}\{W_{k-n}(d_{k-n})+\beta T_{k-n}(d_{k-n})H_n\} \end{aligned} \tag{13}$$

各 狀態 $i\in S$에 대하여 (13)式은

$$H_{n+1}(i)=\underset{a}{\mathrm{Max}}\{W_{k-n}(i,a)+\beta T_{k-n}(j/i,a)H_n(j)\}$$

으로 表示되며 모든 상태를 계산함으로서 各 期間, 各 상태에 대한 最適의 決定을 發見할 수 있다. 時間을 backward horizon으로 부터 측정하고 各 단계에서의 계산은 하나의 決定만이 관여하므로 決定 d_{k-n}의 時間 記號는 없어지게 된다. 따라서 (13)式은 우리가 주로 사용하는 다음과 같은 動的計劃(dynamic programming) 問題로 나타난다.

$$\begin{aligned} H_0&=0\\ H_{n+1}&=\underset{d}{\mathrm{Max}}\ \{W_n(d)+\beta T_n(d)H_n\} \end{aligned} \tag{14}$$

(14)式은 標準 動的計劃形成이며 $H_{n+1}(i)$는 最適의 政策을 使用하였을 때 體系가 狀態 $i\in S$에서 시작하여 $(n+1)$ 期間동안의 總豫想割引收益(expected total discounted profit)로 해석된다. (14)式은 有限의 期間에 대한 最適의 政策을 찾을 때 使用되며 動的 計劃理論을 發展시킨 Bellman의 이름을 붙여 Bellman algorithm이라 흔히 부른다.

2) Howard의 Policy Improvement 技法

앞의 Bellman algorithm이 有限期間에 대한 最適의 政策을 發見하는 方法임에 反하여 Howard의 Policy improvement 기법은 無限期間(Infinite horizon)에 대한 것이다.

만약 費用—收益의 $Wt(j,i,a)$가 모든 狀態와 行動(actions)에 대하여 有限한 값들을 가질 때 豫想收益 $Wt(i,a)$도 有限값을 갖는다. 따라서 政策에 관계없이 時間 k가 無限으로 갈 때 割引된 총예산수익

$$\sum_{t=1}^{k}\beta^{t-1}(P_{t-1}(D),\ Wt(d_t))$$

는 만약 $0\leq\beta<1$이면, 수렴한다. 따라서 各 期間에 대한 政策을 아는 것보다 無限期間에 대한 體系의 政策을 구하는 데 使用된다. 無限期間에 대한 政策은 最適政策이 어떤 規則性

(regularity)을 보여주며, 이 規則性은 계산과정에서 많은 도움을 주기 때문에 더욱 많이 應用된다. 특히 費用—收益構造와 確率이 변하지 않는다면 各 期間에 항상 똑같은 決定을 使用하는 最適의 政策이 반드시 存在한다.

費用 收益과 確率이 時間에 관계없이 一定하다고 假定하면,

$$F(D)=\sum_{t=1}^{\infty}\beta^{t-1}T^{(t-1)}(D)\cdot W(d_t)$$

어떤 特別한 政策 D가 一聯의 無限의 決定들 $d_1, d_2, d_3, \cdots$이라고 하자. 이것으로부터 새로운 政策 D'는 첫번째 기간에 決定 d'를 사용하고 두번째기간 부터는 $d_1, d_2, \cdots$를 사용하는 政策이 될 수 있다. 즉 새로운 政策은 $D'=(d, D)=d, d_1, d_2, \cdots$로 나타내진다.

새로운 政策에 대한 $F(D')$는

$$F(D')=W(d)+\beta T(d)\sum_{t=2}^{\infty}\beta^{t-2}T^{(t-2)}(D)\,W(d_t)$$

로 표시된다.

이것은 $T^{(t)}(D')=T(d),\ T(d_1)\cdots T(d_{t-1})=T(d)\,T^{(t-1)}(D)$를 利用한 것이며 다음과 같이 간단히 쓸 수 있다.

$$F(D')=W(d)+\beta T(d)\cdot F(D) \qquad (15)$$

(15)式은 비록 $F(D)$와 $F(D')$가 여러 出發狀態(starting state)에 대해 無限期間의 政策 D와 D'의 割引된 값을 나타낸 것이지만 근본적으로 Bellman algorithm에서 사용했던 式들과 같은 것이다.

(15)式을 최대화하는 政策을 찾아보자.

$$\operatorname*{Max}_{D}F(D)=\operatorname*{Max}_{(d,D')}\{W(d)+\beta T(d)F(D')\}$$
$$=\operatorname*{Max}_{d}\{W(d)+\beta T(d)(\operatorname*{Max}_{D'}F(D'))\}$$

$\operatorname*{Max}_{D}F(D)$는 또한 $\operatorname*{Max}_{D'}F(D')$이기 때문에 $\operatorname*{Max}_{D}F(D)$와 $\operatorname*{Max}_{D'}F(D')$는 같은 값을 갖는다. 즉

$$H=\operatorname*{Max}_{D}F(D)=\operatorname*{Max}_{D'}F(D')$$

따라서

$$H=\operatorname*{Max}_{d}\{W(d)+\beta T(d)H\} \qquad (16)$$

만약 Vector H가 존재하고 (16)式의 最大값을 주는 最適의 決定 d^*가 있다면, 最適의 決定 d^*는 各 期間에 使用되는 政策이다. 各 期間에 항상 같은 決定 d를 使用하는 stationary 決定을 d^∞라 한자. 이와같이 stationary 決定은 만약 $D=d^\infty$이면, $D=(d, D)$와 같으며 (15)式은 다음과 같이 나타낼 수 있다.

$$F(D)=W(d)+\beta T(d)\cdot F(D) \qquad (17)$$

또는

$$(I-\beta T(d))\ F(D)=W(d)$$

無限期間問題(infinite horizon problem)의 계산은 흔히 Howard algorithm이라 하며, 다음과 같이 계산한다.

i) 任意의 stationary 政策 d^∞를 선택한다.

ii) 政策 d^∞를 使用하여 方程式

$$(I-\beta T(d))\ F(d^\infty)=W(d)$$

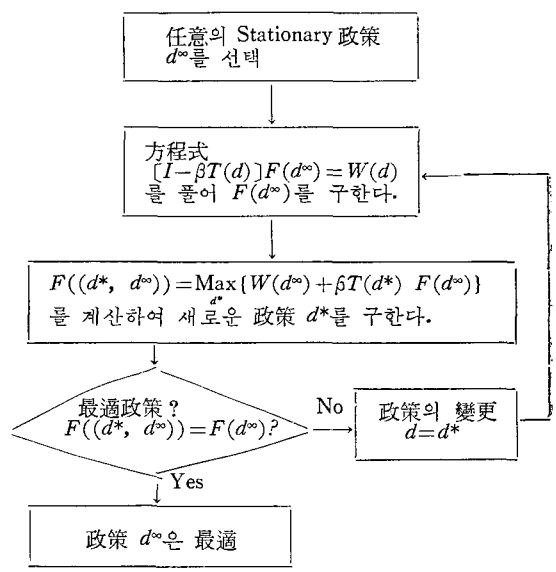
을 푼다.

iii) 위에서 구한 $F(d^\infty)$값을 使用하여 새로운 政策 d'를 구한다.

$$\operatorname*{Max}_{d'}\{W(d')+\beta T(d')\ F(d^\infty)\}$$

iv) 만약 iii)에서 구한 最大값이 $F(d^\infty)$와 같으면 d^∞이 最適의 政策이므로 計算을 中止하고, 그렇지 않으면 새로운 정책을 가지고 ii)로 되돌아가 方程式을 푼다.

그림 1은 Howard algorithm의 계산과정을 표시한 것이다. Howard algorithm에서는 최초의 任意의 Stationary 政策을 택할 때 最適의 政策에 가까운 것을 선택하였다면 여러번 (17)



〈그림 1〉 **Howard algorithm의 Flow chart.**

의 方程式을 풀 必要없이 쉽게 最適의 政策을 發見할 수 있다. 따라서 흔히 最初 Bellman algorithm으로 어느정도 Stationary 政策을 구한 다음에 Howard algorithm으로 無限期間에 대한 政策을 구한다.

III. Sequential decision 理論의 應用

앞에서 살펴본 Sequential decision 理論의 應用問題로 Marketing에서의 廣告政策(advertising policy)에 대한 問題를 分析해 보기로 한다.

販賣高를 올리기 위한 企業들의 商品廣告는 갈수록 熾烈해 가고 있으며 莫大한 廣告費를 支出하고 있다. 氷菓 및 淸凉음료 生產業體인 ×××會社는 다가오는 여름철을 맞아 두가지 廣告方案을 考慮中이다. 方案 1은 新聞, 라디오방송등을 主로하는 소극적인 廣告이며, 方案 2는 T.V. 를 위주로 하는 적극적인 廣告이다. 販賣高의 水準은 여러 단계가 있겠으나 여기에서는 편의상 平均販賣高를 基準으로 低販賣高(狀態 1)와 高販賣高(狀態 2)로 體系의 狀態를 區分한다. Marketing manager는 各 期間에서의 販賣高는 前 期間의 販賣高와 그 期間에 선택한 廣告方案에 의하여 決定된다는 과거의 經驗을 근거로 前 期間의 販賣高와 廣告가 販賣高에 미치는 영향을 條件的確率로 推定하여 移動法則 T를 구하였다.

$T(j|i,a)$를 期間$(t-1)$에 販賣高가 i이고 廣告方案 a를 취하였을때 期間 t에 販賣高가 j가 될 確率이라 하면 推定된 T의 값은 다음과 같다.

$T(1|1,1)=0.9$ $T(2|1,1)=0.1$

$T(1|1,2)=0.4$ $T(2|1,2)=0.6$

$T(1|2,1)=0.5$ $T(2|2,1)=0.5$

$T(1|2,2)=0.4$ $T(2|2,2)=0.6$

따라서

$T(1,1)=(0.9,\ 0.1)$, $T(1,2)=(0.4,\ 0.6)$

$T(2,1)=(0.5,\ 0.5)$, $T(2,2)=(0.4,\ 0.6)$

生產費를 제외한 總收入은 低販賣高때는 5,000萬원이며 高販賣高때는 20,000만원이다. 低販賣高일때, 廣告方案 1에 所要되는 費用은 2,000萬원이고 方案 2에 대한 費用은 10,000萬원이 所要된다. 高販賣高일때는 이미 상당한 廣告效果가 이미 있는 경우이므로 方案 1에 대한 費用은 3,000만원, 그리고 方案 2에 대한 費用은 6,000만원 정도가 所要된다. 또한 販賣高가 變함에 따라 生產率(production rate)도 變함으로 生產率 變化에 따른 費用이 所要된다. 販賣高가 低販賣高에서 高販賣高로 變할 때 所要되는 生產率 變化의 費用은 3,000만원 그리고 高販賣高에서 低販賣高로 變할 때 生產率을 줄이는 데 2,000만원의 費用이 所要된다. 이러한 費用—收益關係로 부터 單位期間에 대한 費用—收益을 계산할 수 있다.

$W(j,i,a)$를 前 期間의 販賣高가 i일때 廣告方案 a를 취하여 販賣高가 j로 될 때 얻어지는 收益이라 한다면

$W(j,i,a)$=(總收入−廣告費−生產率變化費)로 나타난다. 즉

	總收入	廣告費	生產變化費	收益
$W(1,1,1)=$	5,000−	2,000−	0	= 3,000
$W(2,1,1)=$	20,000−	2,000−	3,000	= 15,000
$W(1,1,2)=$	5,000−	10,000−	0	= −5,000
$W(2,1,2)=$	20,000−	10,000−	3,000	= 7,000
$W(1,2,1)=$	5,000−	3,000−	2,000	= 0
$W(2,2,1)=$	20,000−	3,000−	0	= 17,000
$W(1,2,2)=$	5,000−	6,000−	2,000	= −3,000
$W(2,2,2)=$	20,000−	6,000−	0	= 14,000

위에서 구한 收益과 確率을 結合하여 주어진 狀態(販賣高)에서 各 廣告方案에 의하여 얻어진 豫想收益은

$$W(i,a)=\sum_{j=1}^{2} W(j,i,a)\ T(j|i,a),\quad \begin{pmatrix} i=1,2 \\ a=1,2 \end{pmatrix}$$

이며 값들은 다음과 같다.

$W(1,1)=\sum_{j=1}^{2} W(j,1,1)\ T(j|1.1)=4,200$

$W(1,2)=\sum_{j=1}^{2} W(j,1,2)\ T(j|1,2)=2,200$

$W(2,1)=\sum_{j=1}^{2} W(j,2,1)\ T(j|2,1)=8,500$

$W(2,2)=\sum_{j=1}^{2} W(j,2,2)\ T(j|2,2)=7,200$

1. Bellman algorithm의 應用

費用—收益과 確率이 時間에 무관하기 때문에 time horizon을 固定할 必要없이 (14)式을 使用하여 各 期間에 대한 最適의 政策을 Backward procedure로 구하면 된다.

첫번째기간 $n=1$에 대하여

$$H_1=\underset{d}{\text{Max}}\{w(d)\}=\begin{pmatrix}4,200\\8,500\end{pmatrix}$$

따라서 첫번째 기간에서 最適決定 $d_1=(1,1)$

두번째기간 $n=2$에 대하여 各 販賣高에 따라 $H_2=\underset{d}{\text{Max}}\{w(d)+\beta T(d)H_1\}$을 계산하여야 한다. 즉 販賣高 i에 대하여

$$H_2(i)=\underset{a=1,2}{\text{Max}}\{W(i,a)+\beta\sum_{j=1}^{2}T(j|i,a)H_1(j)\},\quad i=1,2$$

低販賣高($i=1$), 方案 $a=1$에 대하여

$$W(1,1)+\beta T(1,1)H_1$$
$$=4,200+0.8(0.9,0.1)\begin{pmatrix}4,200\\8,500\end{pmatrix}=7,904$$

$i=1$, $a=2$에 대하여

$$W(1,2)+\beta T(1,2)H_1$$
$$=2,200+0.8(0.4,06)\begin{pmatrix}4,200\\8,500\end{pmatrix}=7,624$$

따라서 低販賣高일 때는 첫번째 기간과 같이 廣告方案 1을 선택한다.

高販賣高($i=2$), 方案 $a=1$에 대하여

$$W(2,1)+\beta T(2,1)H_1$$
$$=8,500+0.8(0.5,0.5)\begin{pmatrix}4,200\\8,500\end{pmatrix}=13,580$$

$i=2$, $a=2$에 대하여

$$W(2,2)+\beta T(2,2)H_1$$
$$=7,200+0.8(0.4,0.6)\begin{pmatrix}4,200\\8,500\end{pmatrix}=12,624$$

高販賣高일때도 廣告方案 1을 취해야 한다.

期間 $n=2$에 대하여 最適의 政策은

$$d_2=(1,1)$$

세번째期間 $n=3$에 대하여

$H_3(1)=$

Max

$$\begin{Bmatrix}a=1;4,200+0.8(0.9,0.1)\begin{pmatrix}7,904\\13,580\end{pmatrix}=10,976\\ a=2;2,200+0.8(0.4,0.6)\begin{pmatrix}7,904\\13,580\end{pmatrix}=11,248\end{Bmatrix}$$
$$=11,248$$

$H_3(2)=$

Max

$$\begin{Bmatrix}a=1;\ 8,500+0.8(0.5,0.5)\begin{pmatrix}7,904\\13,580\end{pmatrix}=17,094\\ a=1;\ 6,200+0.8(0.4,0.6)\begin{pmatrix}7,904\\13,580\end{pmatrix}=15,248\end{Bmatrix}$$
$$=17,094$$

즉 $d_3=(2,1)$, 低販賣高일 때는 方案 2, 高販賣高일 때는 方案 1을 택한다.

2. Howard algorithm의 應用

任意의 方案 $d'=(1,1)$을 선정한다. 즉 販賣高에 關係없이 方案 1을 취하는 政策에 대하여

$$(\text{I}-\beta T(d'))\ F(d')=W(d') \qquad (18)$$

의 方程式을 푼다. 數値를 (18)에 代入하면,

$$\left[\begin{pmatrix}1&0\\0&1\end{pmatrix}-0.8\begin{pmatrix}0.9&0.1\\0.5&0.5\end{pmatrix}\right]F(d')=$$
$$\begin{pmatrix}0.28&-0.08\\-0.40&0.60\end{pmatrix}\cdot F(d')=\begin{pmatrix}4,200\\8,500\end{pmatrix}$$

위의 方程式을 풀면

$$F(d')=\begin{pmatrix}23,529\\29\text{‘}853\end{pmatrix}$$

다음에는 $\underset{d}{\text{Max}}\{w(d)+\beta T(d)\ F(d')\}$를 계산함으로서 새로운 政策 d를 찾는다.

販賣高 $i=1$에 대하여

Max

$$\begin{Bmatrix}a=1;4,200+0.8(0.9,0.1)\begin{pmatrix}23,529\\29,853\end{pmatrix}=23,529\\ a=2;2,200+0.8(0.4,0.6)\begin{pmatrix}23,529\\29,853\end{pmatrix}=24,059\end{Bmatrix}$$
$$=24,059$$

販賣高 $i=2$에 대하여

Max

$$\begin{Bmatrix}a=1;\ 8,500+0.8(0.5,0.5)\begin{pmatrix}23,529\\29,853\end{pmatrix}=29,853\\ a=2;\ 7,200+0.8(0.4,0.6)\begin{pmatrix}23,529\\29,853\end{pmatrix}=29,059\end{Bmatrix}$$
$$=29,853$$

따라서 첫번째 iteration에서 얻은 政策은 처음에 任意로 선택한 政策 d''보다 向上된 것이다. 즉 $d''=(2,1)$

새로운 政策 d''를 使用하여 다시

$$\{\text{I}-\beta T(d'')\}F(d'')=W(d'')$$

를 푼다.

$$\left[\begin{pmatrix}1&0\\0&1\end{pmatrix}-0.8\begin{pmatrix}0.4&0.6\\0.5&0.5\end{pmatrix}\right]F(d'')=$$

$$\begin{pmatrix}0.68&-0.48\\-0.40&0.60\end{pmatrix}\cdot F(d'')=\begin{pmatrix}2,200\\8,500\end{pmatrix}$$

위의 方程式을 풀면

$$F(d'')=\begin{pmatrix}25,000\\30,833\end{pmatrix}$$

$F(d'')$의 값을 가지고 새로운 政策을 구할 때 앞에서 使用되지 않았던 方案만 檢討해 보면 된다.

販賣高 $i=1$에서 使用되지 않았던 方案 1에 대하여

$$4,200+0.8(0.9,0.1)\begin{pmatrix}25,000\\30,833\end{pmatrix}=24,667$$

이 값은 25,000 보다 적으므로 廣告方案은 必要없다.

販賣高 $i=2$에서 使用되지 않았던 方案 2에 대하여

$$7,200+0.8(0.4,0.6)\begin{pmatrix}25,000\\30,833\end{pmatrix}=30,000$$

이 값 역시 向上된 값이 아니다. 따라서 政策 $d''=(2,1)$, 低販賣高 때는 언제나 方案2 (T.V 廣告)를 택하고 高販賣高 때는 方案 1(新聞, 라디오방송)을 택하는 無限期間에 대한 最適의 政策을 구하게 된다.

IV. 結 論

Sequential decision 理論과 簡單한 應用問題를 通하여 본 바와 같이 Sequential decision 模型은 體系의 動的인 變化와 最適의 政策을 Bellman 의 value iteration 方法과 Howard 의 Policy improvement 技法으로 쉽게 찾아 낼 수 있다. 그러나 體系의 狀態數와 可用한 行動(action)의 數가 크게 증가하는 경우에는 Computer 容量의 不足때문에 計算을 할 수가 없다. 따라서 Sequential decision 問題가 大型化되는 경우에는 計算不能때문에 많은 隘路點을 가지고 있고 大型化問題에는 크게 應用되지 못하고 있다. 이러한 大型化問題의 解決을 위하여 decomposition 理論에 대한 硏究가 한창이다.

體系의 動的인 行態를 分析하는 데 Sequential decision 模型化가 不可能할 때 흔히 Simulation을 使用한다. 비록 實驗을 하는 데 費用은 Sequential decision 이론보다 많이 所要되지만 Simulation 은 Sequential decision 으로 模型化되지 않는 複雜한 模型의 分析에 아주 重要한 技法으로 使用되고 있다.

References

Bellman, R, "A Markovian Decision Process," J. Math Mech. 6, 679-684 (1957)

Bellman, R, *Dynanic Progamming*, Princeton Univ Press Princeton, N.J. (1957)

Blackwell,D, "Discrete Dynamic Programming," Ann. Math. Stat., 33, 719-726 (1962)

Blackwell,D, "Discounted Dynamic Programming," Ann. Math. stat., 35, 226-235 (1965)

Chung, Kai Lai, *Markov Chains with Stationary Transition Probabilities*, Springar, Berlin, 1960

Denardo, E.V. and Fox, B.L, "Multichain Markov Renewal Program," Siam. J. Appl. Math., 16, 468-487 (1968)

Derman, C, *Finite state Markovian Decision Processes*, Academic Press, New York (1970)

Howard,R.A. *Dynamic Programming and Markov Processes*, M.I.T. press, Combridge, Mass., 1960

Kemeny, J.G. and Snell, J.L, *Finite Markov Chains*, Van Nostrand, Princeton, N.J. (1960)

Klein, M, "Inspection-maintenance-replacement Schedules under Markovian Deterioration," Mgt. Sci., 9, 25-32 (1962)

Mann, A.S, "Linear Programming and Sequential Decision," mgt. Sci., 6, 259-267 (1960)

Stranch, R. and Veinott, A.F., *A Property of Sequential Control Processes*, Rand McNally, Chicago, Ill., 1966

Veinott A.F, "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," Ann. Math. Stat 40, 1635-1660 (1969)

Veinott, A.F, "On the Optimality of (S.S) Inventory Policies: New conditions and a New Proof," SIAM. J. 14, 1067-1083 (1966)