DEA によるシミュレーション・ログ集合の分類

A Classification of the Simulation Log Set by DEA

國上真章 [1]　菊地剛正 [2]　寺野隆雄 [3]

Masaaki Kunigami [1], Takamasa Kikuchi [2], and Takao Terano [3]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] 慶應義塾大学
[2] Keio University
[3] 千葉商科大学
[3] Chiba University of Commerce

Abstract: This study proposes a method of using Data Envelopment Analysis (DEA) to classify a set of input/output logs in a social or organizational simulation and compares it with previous methods based on cluster analysis using examples. DEA is used mainly in policy analysis as a method to compare the efficiency based on multiple input and multiple output data. In DEA, the entire data is partitioned by the data that define efficiency corner points, called reference sets. In this study, we propose to use DEA as a classifier of simulation logs due to this property of reference sets. In this paper, we illustrate how DEA allows us to classify a set of simulation logs by their characteristics, using the SIR model of infectious disease spreading with additional variables multiple indicating measures. In addition, the results were compared with the results of the previous log classification using cluster analysis.

1 はじめに

本稿は，社会や組織のシミュレーションの入出力ログ集合の分類に包絡分析法 (Data Envelopment Analysis: DEA)の活用を提案するものである．DEA は多入力・多出力の経営データ間の効率性を相互に比較する手法として，政策分析等で利用されている[1]. DEA による分類では参照集合と呼ばれる効率性の端点を定める経営データによってデータ全体が分割される．本稿では，この参照集合の性質により DEA をシミュレーション・ログの分類器として用いる．

これまでに田中[2]は，組織シミュレーションのログをクラスター分析により分類するとともに，各クラスター毎にその性質を特徴づける意志決定を決定木により抽出する方法を提案している．また菊池[3]は，クラスター分析によりログを類型化し，形式化された仮想ケースとして記述している．

本稿ではシミュレーションの入出力ログをその振舞いに応じて分類するために DEA による方法を提案し，その特徴についてクラスター分析と比較しつつ考察する．

2 包絡分析法（DEA）

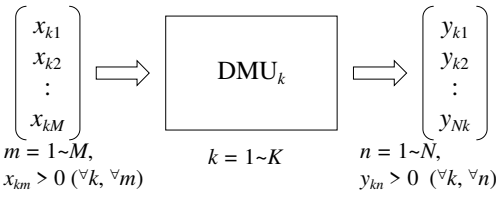
Charnes，Cooper，Rhodes が 1978 年に提案した包絡分析法 (Data Envelopment Analysis: DEA)[4]は，複数の入力と複数の出力をもつ意思決定主体（Decision Making Unit: DMU）の効率性を比較する手法である．DEA では各 DMU について，より最適化された効率性を持つ DMU からなる参照集合（Reference Set）による特徴づけを行うとともに，参照集合をつないで生成される包絡線（Envelopment）により，DMU の相対的な位置づけの比較を可能にする．

2.1 DEA の考え方

DEA の最も基本的な方法である CCR 法[4]の考え方は，次のとおりである．a) 複数の入出力をもつ意志決定主体（DMU）の重み付き効率性，b) 各 DMU の効率の重みの最適化を通した DMU 間の効率性の比較，c) 劣効率的な DMU に対しての最適な効率性を持つ DMU の集まりである参照集合（Reference Set）による特徴づけである．

まず，複数（K 個）の意志決定主体のそれぞれ DMU_k (k = 1~K) が，M 成分の入力 $\mathbf{x}_k = \{x_{km} > 0 \mid m =$

1~M} から N 成分の出力 $\mathbf{y}_k$ {y_{kn} >0 | n = 1~N} を生成するとする．この時，各 DMU$_k$ の効率性Θ_k は，入力 $\mathbf{x}_k$ と出力 $\mathbf{y}_k$ に重み $\boldsymbol{\xi}_k$ ={ξ_{km} >0 (m = 1~M)}, $\boldsymbol{\eta}_k$ (η_{kn} >0 (n = 1~N))をつけて，$\Theta_k \equiv (\boldsymbol{\eta}_k \cdot \mathbf{y}_k) / (\boldsymbol{\xi}_k \cdot \mathbf{x}_k) = \sum_{n,m}\{(\eta_{kn} y_{kn}) / (\xi_{km} x_{km})\}$と定義できる．（図１）



$$\Theta_k \equiv \frac{\boldsymbol{\eta}\cdot\mathbf{y}}{\boldsymbol{\xi}\cdot\mathbf{x}} = \sum_{n=1}^{N}\eta_{kn}y_{kn} \Big/ \sum_{m=1}^{M}\xi_{km}x_{km}$$

図１：意志決定主体 DMU と効率性Θ.

このとき，重み $\boldsymbol{\xi}_k$, $\boldsymbol{\eta}_k$ はいくらでも動かせるので，効率性 Θ_k は変わってしまう．そこで相互に比較するので効率性の上限は１($\Theta_k \le 1$)で共通化した上で，それぞれの DMU$_k$ の効率性 Θ_k を最も有利になるように重み $\boldsymbol{\xi}_k$, $\boldsymbol{\eta}_k$ を最適化する．ただしこの最適化の際，その重み $\boldsymbol{\xi}_k$, $\boldsymbol{\eta}_k$ で他の全ての DMU$_h$ 効率性も 1 を超えない ($\Theta_h \le 1$) ように最適化するのがポイントである．（図２）

$$\max_{\eta,\xi}\ \Theta_k = \sum_{n=1}^{N}\eta_n y_{kn} \Big/ \sum_{m=1}^{M}\xi_m x_{km}$$

DMU$_k$の効率性Θ_k

$$\text{s.t.}\begin{cases} \sum_{n=1}^{N}\eta_n y_{hn} \Big/ \sum_{m=1}^{M}\xi_m x_{hm} \le 1 \quad (h=1\sim K) \\ \eta_n \ge 0 \quad (n=1\sim N) \\ \xi_m \ge 0 \quad (m=1\sim M) \end{cases}$$

その$\boldsymbol{\eta}, \boldsymbol{\xi}$で他のDMUの効率性$\Theta_h$が1を超えないよう制限

重みは全て非負

$$\Rightarrow \xi_{kn} = \xi_n^*, \eta_{kn} = \eta_n^*$$

図２：DMU の効率性Θを求める最適化問題.

もし最適化した重みでもでも効率性 Θ_k =1 にできず他の DMU$_{k'}$ の $\Theta_{k'}$ に劣れば，DMU$_k$ は効率的でないといえる．このとき，DMU$_k$ を最適化した重みにおいてもその効率性 Θ_k を上回り最適な効率性が 1 となる DMU$_{k'}$を DMU$_k$ の参照集合 E_k といい，DMU$_k$ にとっての効率性改善の参考となる．（図３）

$$E_k = \left\{ k' \,\middle|\, \sum_{n=1}^{N}\eta_{kn}y_{k'n} \Big/ \sum_{m=1}^{M}\xi_{km}x_{k'm} = 1 \ \ (1 \le k' \le K) \right\}$$

DMU$_k$を最適化してもΘ$_k$<1のときは，Θ$_k$=1を妨げている他のDMU$_{k'}$（kのための最適化の重みでも効率1となる）が存在する

図３：DMU$_k$の参照集合 E_k.

ここまでに見た DEA の最適化（図２）は，分数計画問題として定式化されているが，実用上は次の等価な線形計画問題に変換して解かれる．（図４）

$$\max_{\eta,\xi}\ \Theta_k = \sum_{n=1}^{N}\eta_n y_{kn}$$

効率性の分母を1に固定する

制約式を等価な不等式に変更

$$\text{s.t.}\begin{cases} \sum_{m=1}^{M}\xi_m x_{km} = 1 \\ \sum_{n=1}^{N}\eta_n y_{k'n} \le \sum_{m=1}^{M}\xi_m x_{k'm} \quad (k'=1\sim K) \\ \eta_n \ge 0 \quad (n=1\sim M) \\ \xi_m \ge 0 \quad (m=1\sim M) \end{cases}$$

$$\Downarrow$$

$$E_k = \left\{ k' \,\middle|\, \sum_{n=1}^{N}\eta_n y_{k'n} = \sum_{m=1}^{M}\xi_m x_{k'm} \ \ (k'=1\sim K) \right\}$$

図４：分数計画問題（図 2）は等価な線形計画問題によって解くことができる.

DEA を用いることの利点の一つが，このように多入力多出力系の効率性を比較する際の重みをどう決めるかという問題が，上記の最適化のプロセスで自動的に解決されることにある.

2.2 DEA による分類

DEA においては，ある DMU$_k$ にとってその参照集合 E_k は，DMU$_k$ にとって改善の参考となる対象であり，効率的な DMU の参照集合は自分自身である．また，入出力変数で張られる空間において DMU の散布図を描くと，効率的な DMU は全 DMU の包絡線となる．また効率的な DMU と原点で分割される領域には，その効率的な DMU を参照集合とする効率的でない DMU が含まれる．よってこれらの DMU 達は同じ参照集合を共有するという性質で分類されていることになる．（図 5）

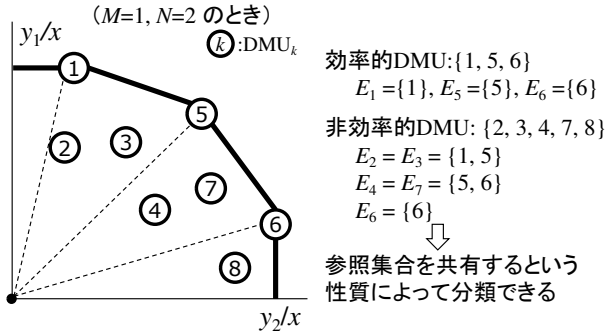


図５：効率的な DMU（参照集合）により包絡線を分割すると非効率な DMU を分類できる.

同じ参照集合を共有する DMU は，効率性の改善について同じ方向（参照集合の元）を共有している．このように，分類群の中において分類群を特徴付ける元（効率的な DMU～参照集合の元）と，その他の

元（非効率的な DMU）の関係が，自動的に与えられることが DEA を用いることの今一つの利点である．

また部分的にしか一致しない参照集合をとるグループ間においては，共通する部分とその多少によってグループ間の類似性を考察することができる．これも分類器としての DEA を見たときの利点である．

3 シミュレーションへの適用

簡単なシミュレーションを用いて，その入出力データを DEA により分類してみる．シミュレーションモデルとしては，感染症伝播に関する SIR モデル[5]を用いる．ただし SIR モデルは，独立な自由度が２しかないため，出力の自由度とともに政策的な入力変数を追加して DEA を適用する．

3.1 SIR モデルの拡張

本稿で例題として用いる拡張した SIR モデルは以下の状態遷移（図６）を持つ差分方程式系（式(1)～(3)）である．

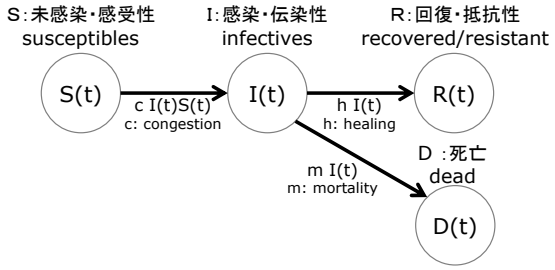


図６：拡張された SIR モデルの状態遷移と
差分方程式系および制御入力

$$\begin{cases} S(t+1)=S(t)-c(t)S(t)I(t), \\ I(t+1)=I(t)+c(t)S(t)I(t)-h(t)I(t)-m(t)I(t), \\ R(t+1)=R(t)+h(t)I(t), \\ D(t+1)=D(t)+m(t)I(t). \end{cases} \quad \cdots\cdots(1)$$

$$\begin{cases} P(t)=S(t)+I(t)+R(t), \quad \text{population alive}, \\ P(0)=P_0, \quad \text{initial population}, \\ S(0)=P_0-\varepsilon, \quad I(0)=\varepsilon,, \quad R(0)=0, \quad D(0)=0. \end{cases} \quad \cdots\cdots(2)$$

$$\begin{cases} c(t)=\dfrac{c_0}{1+\alpha_u \cdot u(t)}, & :\text{contagion}, \\ h(t)=h_0(1+\alpha_v \cdot v(t)), & :\text{heeling}, \\ m(t)=\dfrac{m_0}{1+\alpha_w \cdot w(t)}, & :\text{mortality}. \end{cases} \quad \begin{cases} u(t),v(t),w(t)\ge 0 \\ u(t)+v(t)+w(t)\le 1 \end{cases} \cdots(3)$$

式(1)は状態変数の時間変化を，式(2)は従属変数と状態変数の初期値を，式(3)は制御変数による係数の時間変化を表す．

状態変数は，未感染の感受性人口を $S(t)$，感染者人口を $I(t)$，回復した非感受性人口を $R(t)$ とし，さらに死亡人口を $D(t)$として追加した，従属変数 P(t)は生存者人口である．また感染率 c，治癒率 h，死亡率 m を，制御変数 $u(t), v(t), w(t)$ によって変化させられる変数とした．ここで正定数 c_0, h_0, m_0 は制御入力がない時の $c(t), h(t), m(t)$ の値を示し，正定数 $\alpha_u, \alpha_v, \alpha_w$ はそれぞれ $u(t), v(t), w(t)$の感度である．

制御入力 $u(t), v(t), w(t)$について，$u(t)$は感染率 $c(t)$を抑制する施策を，$v(t)$は治癒率 $h(t)$を向上させる施策を，$w(t)$は死亡率 $m(t)$を抑制する施策をそれぞれ表す．これらの費用と効果はそれぞれ異なるが費用で規格化することで一般性を失わずにその違いを感度 $\alpha_u, \alpha_v, \alpha_w$ のみで表し，費用制約は $u(t) + v(t) + w(t) \le 1$ としている．

3.2 DEA による結果の分類

前節で導入した拡張された SIR モデルを例として用いて，シミュレーションの入出力データの集合を DEA によって分類する．共通する初期値，係数等は表１のとおりである．

Initial Conditions	$S(0)$: Susceptibles	$I(0)=\varepsilon$: Infectives	$R(0)$: Recovered	$D(0)$: Dead
	9990	10	0	0
	$P(0)$: Population Alive =10000			
Coefficients	c_0 Contagion	h_0 Healing	m_0 Mortality	
	0.00002	0.01	0.01	
Control Sensitivities	α_u	α_v	α_w	
	10	15	20	

表１：シミュレーションの初期値および係数

次に，シミュレーション実行(RUN)設定は，表２に示す 10 通りの制御入力の割付に従った．

		Stage-0			Stage-1			Stage-2			Stage-3		
		t = 0 ~ 29			t = 30 ~ 89			t = 90 ~ 179			t = 180 ~ 500		
		$u(t)$	$v(t)$	$w(t)$	$u(t)$	$v(t)$	$w(t)$	$u(t)$	$v(t)$	$w(t)$	$u(t)$	$v(t)$	$w(t)$
RUN # (DMU #)	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.3	0.3
	1				1.0	0.0	0.0	1.0	0.0	0.0			
	2				1.0	0.0	0.0	0.0	1.0	0.0			
	3				1.0	0.0	0.0	0.0	0.0	1.0			
	4				0.0	1.0	0.0	1.0	0.0	0.0			
	5				0.0	1.0	0.0	0.0	1.0	0.0			
	6				0.0	1.0	0.0	0.0	0.0	1.0			
	7				0.0	0.0	1.0	1.0	0.0	0.0			
	8				0.0	0.0	1.0	0.0	1.0	0.0			
	9				0.0	0.0	1.0	0.0	0.0	1.0			

表２：シミュレーションの実行設定．

上記の設定から得られた実行結果の要約値は表３のとおりである．要約値を用いたのは，本例では実行数が少ないためと，DEA の性質から出力変数は正の望大特性であることが必要なためである．

		Input:		OutPut:		
		Cs	Cm	BtmActP	AvgActP	AlvPE
		Social Cost:	Medical Cost:	Bottom Active Population:	Average Active Population:	Alive Poplation at the end:
		接触・感染防止の社会的コスト: $\mathrm{Avg}[P(t)\cdot\alpha_u \cdot u(t)]$	感染者の回復と救命コスト: $\mathrm{Avg}[I(t)\{\alpha_v \cdot v(t) + \alpha_w \cdot w(t)\}]$	活動可能な人数が最も落ち込んだ時の値: $\mathrm{Min}[P(t)-I(t)]$	活動可能な人数の平均値: $\mathrm{Avg}[P(t)-I(t)]$	最終的に生き残った人の数: $\mathrm{Min}[P(t)]$
Run # (DMU #)	0	13,574	231	1,915	5,112	5,296
	1	49,693	340	8,176	8,590	8,638
	2	34,737	617	8,371	9,011	9,072
	3	34,481	19,220	2,514	7,974	8,967
	4	41,879	912	8,650	9,560	9,603
	5	24,652	919	8,650	9,575	9,619
	6	24,328	6,776	6,505	9,147	9,488
	7	33,203	15,244	1,906	6,589	7,235
	8	24,088	15,932	1,906	8,663	9,399
	9	24,418	29,615	1,906	8,082	9,524

表３：シミュレーションの実行結果
（入出力要約値）

表４は，シミュレーションの入出力要約値のデータ（表３）から DEA（CCR 法）によって求めた，最適な効率性Θ（DEA Score），参照集合（Reference Set），入出力の要約値への最適化された重み（Input Weight, Output Weight）である．なお CCR 法の計算については，並木[6]による Python コードを用いた．

DEA_CCR		DEA Score Θ	Reference Set	Input Weights		Output Weights		
				Cs	Cm	BtmActP	AvgActP	AlvPE
Run # (DMU #)	0	1.0	{0, 1, 2}	0.4	22.4	1.8	0	1.2
	1	1.0	{1}	0	29.4	1.2	0	0
	2	1.0	{1, 2}	0.1	8.4	1.2	0	0
	3	0.6665	{8, 5}	0.3	0	0	0	0.7
	4	0.7903	{0, 2, 5}	0.1	5.3	0.5	0.4	0
	5	1.0	{5}	0.4	0	1.2	0	0
	6	0.9995	{8, 5}	0.4	0	0	0	1.1
	7	0.5584	{8, 5}	0.3	0	0	0	0.8
	8	1.0	{8}	0.4	0	0	0	1.1
	9	0.9997	{8}	0.4	0	0	0	1.0

表４ :DEA による表３の入出力要約値の分析結果

DEA の結果（表４）によると，効率的な実行結果として，{Run 0, 1, 2, 5, 8}が得られた．一方，非効率な実行結果としては，{Run 3, 6, 7}が{Run 5, 8}を参照集合として共有する分類群となっており，入出力要約値の重みづけも互いに近い傾向を示している．

また効率的な実行結果である，{Run 0, 1, 2, 5, 8}については，Run 8 と 5 にはこれらを参照集合として共有するグループがあり，Run 0, 1, 2 は互いを参照集合としているなどそれぞれの間で違いが少ないことを示している．反面，Run 8 と Run 0, 1 の間にはこれらを共有する非効率集合が無いなど，これらの間で違いが大きいことを示している．

これらの結果を模式的にまとめたのが図 7 である．効率的な実行結果（参照集合）の共有によって，実行結果がグループ化されている，これによりグループ内での効率性の改善の参考となるデータが明らかになっている．また最適な入出力要約値の重みについて，グループの特徴づけがされている．さらに，参照集合の共有の有無によってグループ間の近さも読み取ることができる．

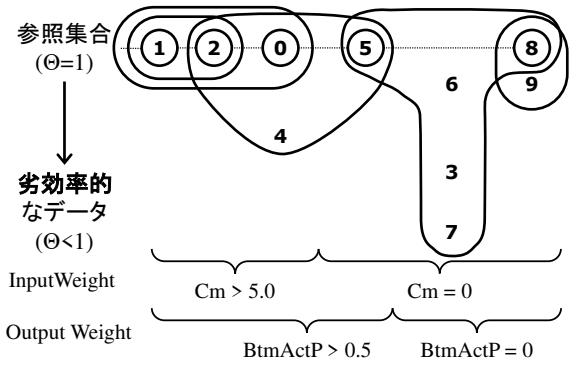


図７：DEA による入出力要約値のグループ化

3.3 クラスター分析による分類

次に，同じ入出力要約値（2 入力 3 出力）について，クラスター分析による分類を行う．分類は階層的クラスタリングの Ward 法により，データ間の距離は Euclid 距離を用いた．

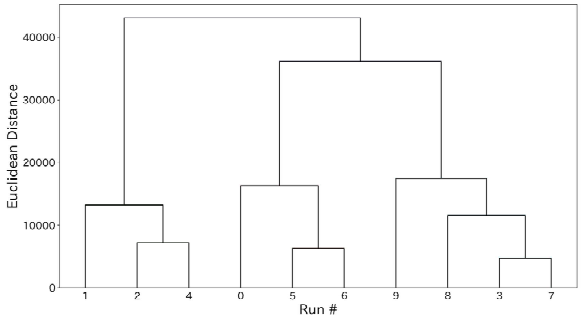


図８：入出力要約値のクラスタリング

クラスタリングによる，入出力要約値の類似性（ベクトルとしての距離の近さ）による樹形図は図８のとおりである．

4 考察

これまでに見たように DEA による分類とクラスター分析では，同じデータに対しても異なった分類を示す. ここでは DEA とクラスター分析のアプローチについて分類器としての特徴という視点でから追加的に考察する．

まず基本的なアプローチについて，DEA ではグループの分類と特徴づけは，参照集合によって行われる．参照集合はグループの端点であり，グループ内でも特異な元である．このためクラスター分析のようなグループ内の平均あるいは重心付近の元という意味は持たないが，グループ内の元の効率性の方向性でのトップという意味は明確である．

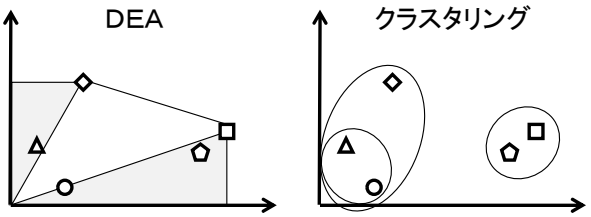


図８：クラスタリングが互いに近いものをまとめるのに対し DEA は端点でグループ化する.

DEA による分類においては，入出力の重みは自動的に最適化される．このためクラスター分析でのデータの重みをどう決めるかという問題は，DEA 任せにできる．一方 DEA には，分類に特化した方法ではないための使い勝手の悪さも存在する．例えば，クラスター分析では，生成するクラスターの階層数を選ぶことができる．これ対し DEA では，グループの数は DEA 任せであり，一つのグループしか生成されないことや，すべてのグループがひとつの参照集合しか含まないということもあり得る．また DEA では，入力は正の望小特性，出力は正の望大特性である必要があり，負の値や望大特性と望小特性が入り交ざっていると事前にデータの変換処理が必要になる．

次に，データの追加に対する安定性について比較してみる．まず，グループの内側に非効率なデータが追加される場合，DEA によるグループ分けは影響を受けない．一方クラスター分析ではクラスター間に中間的なデータが追加されるとクラスタリングが大きく影響される可能性がある．（図９上）

他方，新たな包絡線の外側に位置するようなデータが追加された場合，DEA によるグループ分けは大きな影響を受ける．一方，クラスター分析では大きな外れ値は独立したクラスターとなり，その他のクラスタリングには影響しない．（図９下）

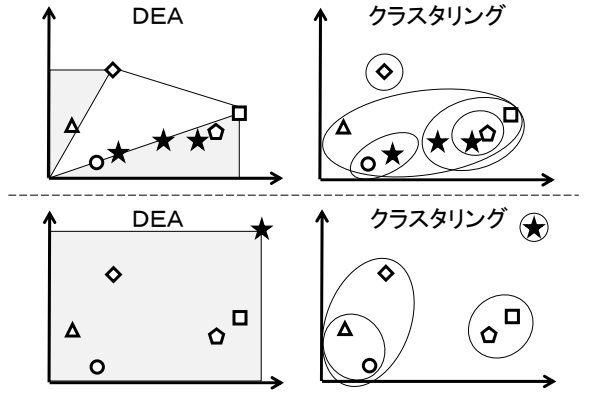


図９：新しいデータの追加に対する安定性：図８のグループの内部に追加（★）された場合（上）とグループの外側に追加（★）された場合（下）

このように DEA による分類は，クラスター分析に対して相反する性質を有する．このことから DEA による分類は，クラスター分析を超えるものというよりは，互いに補う様に使うことでより分析の幅を広げられる可能性を持つと考えられる．

5 まとめ

本稿は，シミュレーション入出力ログ集合の分類法としての DEA の活用を提案し，シミュレーションによる例題と，クラスター分析との比較を考察した．DEA による分類は，クラスタリングと相互補完的な分類器となると考えられる．ログの本格的な分類器として活用するためには，Window 分析等の時系列の DEA も必要であり，今後検討を深めていきたい．

謝辞

本稿の着想の多くは，竹内俊彦先生の解説[7]に拠っている，この場で深く感謝を表したい．また，本研究の一部について，公益財団法人科学技術融合振興財団の調査研究助成を受けている．

参考文献：

[1] Cooper, WW., Seiford, LM., Tone, K.: Data Envelopment Analysis, A Comprehensive Text A Comprehensive Text with Models, Applications, References and DEA-Solver Software (Second Edition), Springer, (2007).

[2] 田中祐史, 國上真章, 寺野隆雄: エージェントシミュレーションにおけるログクラスターの系統的分析からわかること，シミュレーション＆ゲーミング，Vol.27, No.1, pp. 31-41, (2017)
DOI https://doi.org/10.32165/jasag.27.1_31

[3] 菊地剛正, 國上真章, 高橋大志, 鳥山正博, 寺野隆雄 ビジネスケースの形式的記述のためのシミュレーション結果の類型化手法, 経営情報学会論文誌 研究ノート Vol.29, No.3, (to be published 2020)

[4] Charnes, A., Cooper, WW., Rhodes, E.: Measuring Efficiency of Decision Making Units, European Journal of Operational Research, Vol.2 , pp.429-444, (1978).

[5] Kermack, WO., McKendrick, AG.: A Contribution to the Mathematical Theory of Epidemics, Proceedings of the Royal Society, vol.115A, pp.700‑721 (1927),
DOI https://doi.org/10.1098/rspa.1927.0118,
(reprinted Bulletin of Mathematical Biology vol.53, pp.33-55 (1991))

[6] 並木誠: Python による数理最適化入門，朝倉書店，pp.70-71, (2018)

[7] 竹内俊彦：DEA 入門・泉恵女学園物語，http://blog.livedoor.jp/lionfan/archives/52682140.html, (2020.8 閲覧)