

Construction of a Comprehensive Analysis Platform for Typology and its Application

Shinya MIZUNO ^{†1}, Yoshikazu FUJISAWA ^{†2} and Naokazu YAMAKI ^{†3}

Abstract

There is an ongoing effort to consolidate existing elements in various fields such as government, medicine, and education. At the same time, society is facing considerable challenges, such as a projected aging of the population, a decline in the younger population, and a shortage of workers. To respond to these issues, it is necessary to classify the features of the objects in each field and to propose efficient improvements. In this research, we propose a structure and flow for a classification analysis foundation, as well as an environment in which classifications can be performed on any scale. A data envelopment analysis (DEA) is used for the classification in order to clarify the features of each element. We then use the analysis foundation to classify medical and educational institutions in order to investigate the type of categories produced. Various approaches have been used for classification. Methods used in cluster analyses include the farthest neighbor method and the k-means method. However, because these methods use average values, it is possible that the features of individual elements may remain unclear. Moreover, classification foundations include a geographically specialized foundation, which does not lend itself to general use. Furthermore, many classification analyses restrict the number of objects, casting doubt on their ability to cope with large-scale classification analyses. In this research, we propose an analysis foundation that is capable of general-purpose, large-scale classifications. The classification analysis foundation is structured to enable the classification of many objects, including medical institutions, educational institutions, and patient analyses. Here, APIs are used to provide the subject data for the classifications, and a DEA is used for the calculations to clarify the features of the objects.

Key words: typology, DEA, CCR, linear programming, analysis platform

^{†1} Shizuoka Institute of Science and Technology

^{†2} Miyagi University

^{†3} Shizuoka University

Received: January 10, 2016

Accepted: March 2, 2017

総合的類型化解析基盤の構築とその応用

水野 信也 ^{†1}, 藤澤 由和 ^{†2}, 八巻 直一 ^{†3}

現在、行政、医療、学校など様々な分野で既存要素の統廃合が進んでいる。また今後予想されるさらなる高齢化社会、若年者人口減少、労働者の不足などを大きな社会問題として抱えている。そのような問題に対応するために、各分野での類型化を実施して対象毎の特徴を明らかにし、効率的な改善が必要である。またビッグデータ環境が広がり、多くのデータを取り込み、解析を実施するためには、整備された解析環境が必要である。またオープンデータや WebAPI を利用することで、新しい要素を取り入れた解析が可能になる。本研究ではこのようなビッグデータ環境の中で類型化を実施するための類型化解析基盤の構造およびフローを整理し、任意の規模で類型化を実施できる環境を提案する。類型化には DEA を用い、各要素の特徴を明確にするとともに改善指標を示し、結果を可視化する。その類型化解析基盤を用いて、実際にある医療機関の類型化を実施し、特徴を持った分類と現実的な時間内での計算を確認する。

キーワード： 類型化, DEA, 線形計画法, 解析基盤, ビッグデータ, オープンデータ, 可視化

1 はじめに

現在、行政、医療、学校など様々な分野で既存要素の統廃合が進んでいる。また今後予想されるさらなる高齢化社会、若年層の人口減少、労働者の不足など、社会は大きな問題を抱えている。そのような問題に対応するために、各分野での類型化を実施して対象毎の特徴を明らかにし、効率的な改善を提起することが必要である。またビッグデータ解析が盛んに叫ばれている現在、企業が所有するデータを使い経営効率を完全とする試みが見られるが、専門的な知識や技術および解析基盤の不足からなかなか解析が進まないことも多い。

本研究ではこのようなビッグデータ社会に対応するため、類型化が容易に実施可能な類型化解析基盤の構造およびフローを提案する。ビッグデータ解析で注目される要素の一つとして、特徴抽出がある。特徴抽出には、広くデータマイニング技術 [1] を用いられていたが、経営層にも理論面でわかりやすく、特徴を明示しやすい DEA(Data Envelopment Analysis) をいる。ビッグデータを DEA によって解析するためには明確なデータベース構造が必要になる。このデータベース構造は汎用的に用いることができ、データの用意、計算の実施、結果の可視化が可能になる。この類型化解析基盤を用いて実際にある医療機関や教育機関の類型化を実施し、どのような分類が出来るか検証していく。

類型化には今まで様々なアプローチがされてきた。クラスタ分析に用いられる手法として最速隣法 [2] や k-means 法 [3] がある。これらは基本的には平均値を

利用しており、各要素の特色が明確になっていない可能性がある。また類型化基盤では地理的に特化した基盤 [4] があるが汎用的な類型化解析基盤になっていない。また類型化解析の中では対象数が限られているものも多く [5][6]、大規模な類型化解析が可能か疑問が残る。また DEA を用いた大規模計算例では、3,353 箇所項目数 3 を用いて約 8 時間程度で計算した例がある [7]。ビッグデータ環境に対応するためには、使用する対象数の増加や計算時間短縮をする必要があり、またその結果が効率的に利用できるようなデータベース等への保存環境が重要である。

DEA の計算環境は分散型コンピューティング環境が提案されている [8] がデータベース構造や可視化には触れられておらず、ビッグデータ環境にそのまま対応するのは難しい。また計算スケールに着目すると、大規模な CCR モデルを実用時間で解く試みが示されているが [9]、結果の保存環境及び保存に費やす時間、また参照集合の算出などには言及されておらず、実在データでの計算例の提示は無い。また現在はクラウド・コンピューティング環境が利用されており、サーバのスケールアップやスケールアウトが可能である [10]。またオープンデータの利用や様々な API の利用も広がり、それらも取り入れた環境が必要である。またビッグデータ解析においては、対象分野における知見を持った専門家とディスカッションする場が必要であり、ここではデータの可視化が不可欠である。そこで本研究では、ビッグデータ環境に対応可能な汎用的な類型化解析基盤として、データベース及びシステム構成を明示し、その結果を可視化する手法を示す。またこの類型化解析基盤を用いた場合の計算時間基準を示し、実用的に対応出来る事を示す。さらにオープンデータや WebAPI を利用し、外部からの情報を用いて入出力項

^{†1} 静岡理工科大学

^{†2} 宮城大学

^{†3} 静岡大学

受付：2016 年 1 月 10 日、再受付 (2 回)

受理：2017 年 3 月 2 日

目を構成する．今回類型化例として，全国約 7000 病院を対象に類型化を実施し，参照集合やその従属集合及びそのウエイトを算出する．またその結果を可視化し，改善のためのディスカッションに活用する．本研究は OR の分野で研究されてきた DEA をビッグデータ環境に適用する事例と考えている．

類型化の解析基盤の全体像は図 1 のように，医療機関，教育機関または患者分析など多くの対象を類型化出来るように構築していく．また類型化対象のデータを補うために各種 API を用いていく．また実際の計算には DEA を用いて特徴を明確にし，類型化を行う．

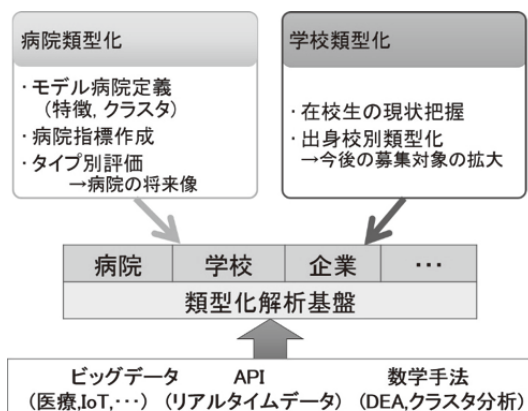


図 1 類型化解析基盤の全体像

2 類型化解析基盤の構成

2.1 類型化解析フロー

類型化を実施するには図 2 にあるように，次のようなフローになる．

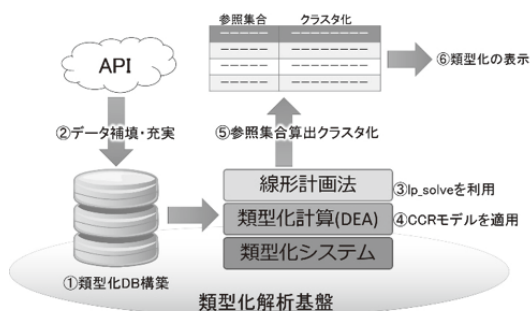


図 2 類型化解析のフロー

- (1) 類型化データベースの構築．
- (2) 類型化データベースに対して，各種 API を用いてデータの補填，充実を図る．
- (3) 類型化システムに対し，線形計画法を計算可能な環境を設定．今回は lp_solve[11] を利用．
- (4) DEA の CCR モデルを用いて各対象の評価を実施．
- (5) DEA 評価の集計を実施し，参照集合を算出しクラスタ毎に分類．
- (6) 類型化の表示．

2.2 類型化データベースの作成と API の利用

類型化データベースは基本的に図 3 のような形で作成する．このような形であれば，どのような対象であっても同様に類型化計算が可能である．まず情報テーブル群として DMU(Decision Making Unit) テーブルと DMU 関連テーブルで類型化に必要な情報を用意する．そして計算テーブル群として，組合せテーブル，結果テーブル，結果関連テーブルで全ての DMU に対しての結果を保存する．

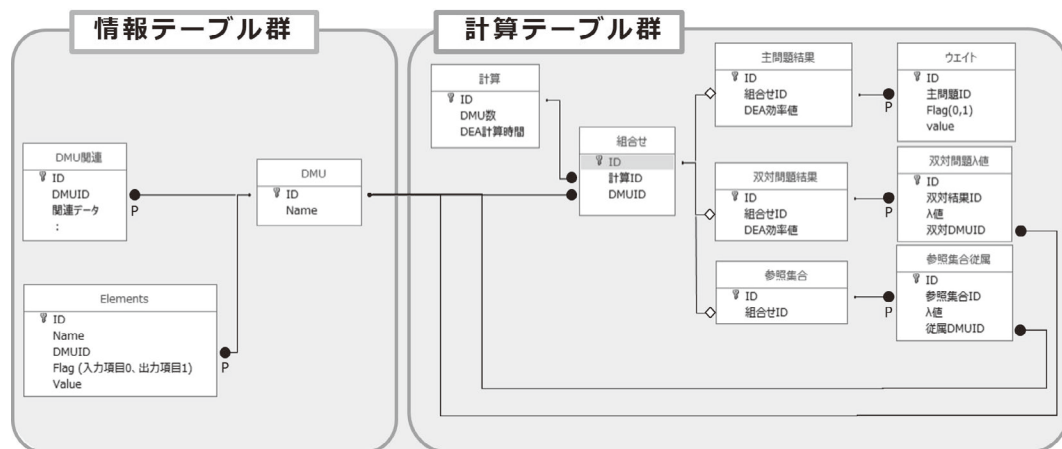


図 3 類型化データベース構造

また情報テーブルに各種 API を利用してデータを入力する場合のフローは図 4 のようになる。情報格納クラスがデータベースと連携し、API から得た情報を格納する。また API からデータを取得するために、API 取得クラスのインスタンスを生成し、API から情報を取得出来る関数を呼び出す。このような構造で幅広い API に対応出来る。

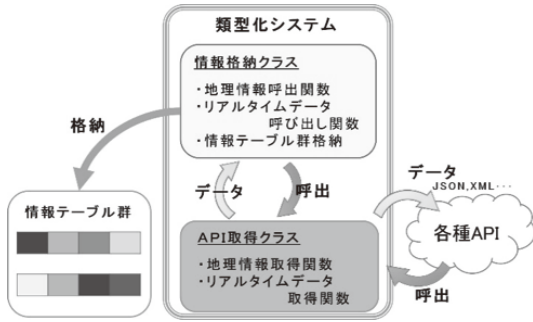


図4 APIから情報を取得するフロー

2.3 DEA の利用

今回提案する類型化解析基盤では、DEA[12][13][14]を用いて多要素を分析し、対象の特徴を活かしながら評価を実施していく。特に線形計画問題として捉えることが出来るCCRモデル[15][16]を利用していく。

2.3.1 DEA のCCRモデル定式化

計算対象となるDMUを $DMU_j (j = 1, 2, \dots, n)$ とすると、 DMU_j の i 番目入力項目データを $x_{ij} (i = 1, 2, \dots, m)$ 、その入力データのDEAウェイトを ν_{ij} とする。同様に r 番目の出力項目データを $y_{rj} (r = 1, 2, \dots, s)$ 、その出力データのDEAウェイトを μ_{rj} とする。 $DMU_k (k = 1, 2, \dots, n)$ に対してDEA効率値を計算する場合の主問題と双対問題の定式化は次のようになる。

[DMU_k に対する主問題]

$$\begin{aligned} \text{目的関数 } & \text{Max} \left(\sum_{r=1}^s \mu_{rk} y_{rk} \right) \\ \text{制約条件} & - \sum_{i=1}^m \nu_{ik} x_{ij} + \sum_{r=1}^s \mu_{rk} y_{rj} \leq 0 \\ & (j = 1, 2, \dots, n) \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^m \nu_{ik} x_{ik} &= 1 \\ \nu_{ik}, \mu_{rk} &\geq 0 (i = 1, 2, \dots, m), (r = 1, 2, \dots, s) \end{aligned}$$

[DMU_k に対する双対問題]

目的関数 $\text{Min } \theta_k$

制約条件

$$\begin{aligned} - \sum_{j=1}^n x_{ij} \lambda_{jk} + \theta_k x_{ik} &\geq 0 \\ (i = 1, 2, \dots, m) \end{aligned}$$

$$\begin{aligned} \sum_{j=1}^n y_{rj} \lambda_{jk} &\geq y_{rk} \quad (r = 1, 2, \dots, s) \\ \lambda_{jk} &\geq 0, (j = 1, 2, \dots, n) \end{aligned}$$

上記は DMU_k についてのDEA効率値計算式だが、これを $k = 1, 2, \dots, n$ まで全てのDMUに対し実施する。

2.3.2 lp_solve の利用

DEAをCCRモデルで扱うことで、線形計画法として計算を実施する。今回はlp_solveを採用している。採用した理由はlp_solveの次のような特徴による。

- ・制約条件数が1万件を超えるような計算が可能
- ・オープンソースであり、学術的なチェックが可能
- ・多くの言語のAPIをサポートしており、多くの環境から利用可能

lp_solveを利用して、DEAの全体計算を行う過程は図5、図6、図7のようになる。

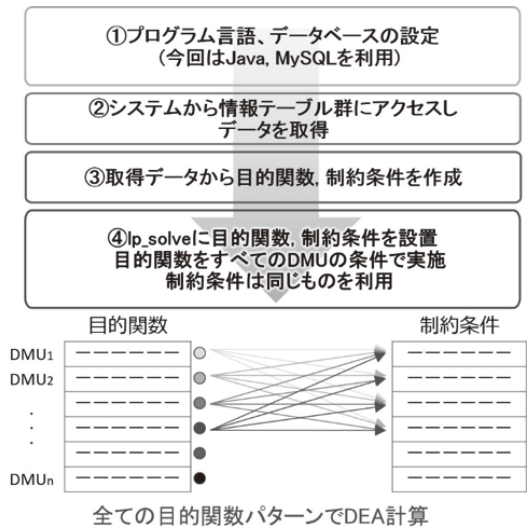


図5 類型化解析フロー概要

2.3.3 参照集合の算出とDMUの分類

今回の類型化解析基盤では、大規模な類型化を想定している。DEAのDMUの分類には幾つか提案されているが[17][18]、今回は明確なプロセスで表現され

利用API	用途
setMaxim()	最大値として設定
strsetObjFn(目的関数オブジェクト)	目的関数設定
Solve()	線形計画法を解く
strAddConstraint(制約条件のオブジェクト, 条件式, 右辺)	制約条件の追加
getObjective()	目的関数値
GetPtrVariables()	係数の取得

図 6 DEA 計算: lp_solve の主な利用関数

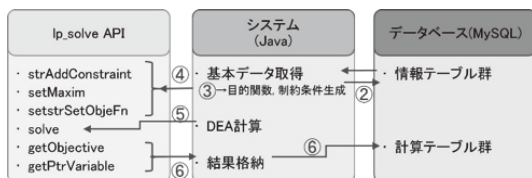


図 7 類型化プログラムフロー

る双対問題のウェイトを用いた方法で DMU の分類を行う [19][20].

- (1) 主問題及び双対問題の目的関数値を計算し, 目的関数値 = 1 となる DMU を算出し, 参照集合を形成する. 対象 DMU を参照集合テーブルに格納する.
- (2) (1) において目的関数値 < 1 の DMU_k に対し, λ_{jk} , ($j = 1, 2, \dots, n$) が $\lambda_{jk} > 0$ となる時, DMU_k は DMU_j を参照しているとし, 参照集合従属テーブルに関係を格納する.

3 数値計算例

提案した類型化解析基盤を用いて数値計算を実施する. 今回は日本全国の病院を対象とした病院類型化について紹介する. DEA の入力項目と出力項目については, 次のものを選択した.

- 入力項目 (医師数, 最寄り駅までの距離)
- 出力項目 (病床数, 診療科数)

ただし医師数は常勤数とする. また最寄り駅までの距離は株式会社ハートレイルズの最寄り駅情報取得 API[21] を利用している. 上記の項目をすべて取得できた病院は 7713 病院あり, 今回はこれらの病院を対象として計算する.

3.1 静岡県内病院での計算例

本研究における類型化解析基盤では全国の病院を対象に計算可能であるが, 情報を明確にするため, 地域を限定した形で計算を行う. 静岡県内病院ですべての項目を取得できた病院は 158 件であった. 表 1 はその情報の一部である. 表 2 は DEA 計算結果から類型化したものである. 表中の第 2 列から第 5 列は上下 2 段になっており, 上部は用いたデータ, 下部のカッコ内数値は全 DMU 内の順位となっている. 第 2 列及び第

3 列は入力項目であり, 値を昇順に並べた順位となっている. 第 4 及び第 5 列は出力項目であり, 値を降順に並べた順位となっている. この中で 22029 は医師数が最小であり, 出力 2 項目は比較的良好な数値となっている. また参照数も 102 と最大となっている. 医師数が最小である病院に 22041 もあるが, こちらは最寄り駅への距離も近い特徴がみられ, 22041 を参照している病院数は 6 と少ない. また最寄り駅に近い病院として 22180, 22032, 22055 及び 22061 がある. 22180 は最も最寄り駅に近く, 22032 は最寄り駅に近いことに加え, 出力 2 項目の数値も良い. それぞれが何らかの特徴を持ち, 参照集合となっている. また今回は入力指向評価であることから, 小中規模の病院の DEA 効率値が高くなっている.

表 1 静岡県内病院の入力及び出力項目 (15 病院)

病院コード	医師数 (常勤)	最寄り駅距離 (m)	病床数	診療科数
22003	24	3530	410	9
22005	7	2330	284	6
22007	77	1910	450	24
22009	7	9710	175	8
22010	387	1860	613	18
22011	2	9050	50	4
22012	45	1750	350	20
22013	18	4050	350	6
22014	155	3670	200	17
22015	245	2850	720	24
22016	158	6210	500	21
22018	25	440	320	18
22019	28	3390	270	12
22020	149	2990	606	23
22021	21	540	200	18

3.2 類型化解析結果の可視化

表 2 の類型化結果より参照集合とそれを参照する従属集合が明らかになった. 得られた結果をディスカッションするために可視化を行う必要がある. DEA の特徴は参照すべき DMU とその比率が明らかになり, その指標に基づいて改善することができる. 表 3 は前節で得られた病院類型化の結果から, ある非参照集合に対応する参照集合の組み合わせとその比率をまとめた一部である. 今回の結果からは最大で 3 個の参照集合を参照しており, 3 個の参照集合組み合わせは 7 通りであった. また 2 個の参照集合組み合わせは 5 通りであり, 1 個だけで参照されている参照集合も 2 通りあった. 今回の可視化は 3 次元表示に対応し, 表示グラフを多角度から見られる gnuplot[22] を利用している. データベースから CSV でエクスポートしたデータを gnuplot で読み込み表示を行っている.

表2 病院類型化の参照集合一覧

参照集合	医師数	最寄駅距離	病床数	診療科数	クラス	特徴
22029	2 (1)	3170 (112)	400 (22)	18 (21)	22003, 22005, 22009, 22011, 22013, 22019, 22027, 22028, 22042, 22046, 22047, 22051, 22052, 22054, 22058, 22059, 22064, 22066, 22067, 22068, 22072, 22074, 22075, 22076, 22077, 22079, 22080, 22084, 22085, 22086, 22087, 22088, 22091, 22092, 22093, 22094, 22095, 22098, 22100, 22104, 22106, 22108, 22109, 22110, 22111, 22112, 22116, 22117, 22119, 22120, 22121, 22123, 22125, 22126, 22128, 22129, 22130, 22131, 22132, 22139, 22142, 22145, 22147, 22148, 22149, 22150, 22151, 22152, 22153, 22154, 22156, 22157, 22158, 22162, 22163, 22164, 22165, 22166, 22168, 22169, 22171, 22172, 22174, 22175, 22176, 22177, 22184, 22185, 22188, 22189, 22190, 22191, 22192, 22193, 22197, 22198, 22199, 22200, 22202, 22203, 22204, 22206 (全 102 力所)	従属集合数が最大数 医師数が最小 出力項目の2項目の評価は高い
22032	50 (137)	140 (2)	350 (25)	13 (33)	22010, 22023, 22037, 22043, 22044, 22060, 22137, 22201 (全 8 力所)	最寄り駅まで近い 両出力項目の評価は比較的高い
22036	5 (34)	330 (12)	99 (126)	8 (56)	22003, 22009, 22012, 22014, 22016, 22019, 22021, 22022, 22025, 22027, 22028, 22033, 22035, 22038, 22042, 22046, 22047, 22050, 22051, 22052, 22054, 22059, 22064, 22067, 22072, 22076, 22081, 22089, 22092, 22102, 22107, 22111, 22114, 22119, 22122, 22124, 22125, 22132, 22143, 22144, 22145, 22148, 22149, 22152, 22154, 22163, 22165, 22170, 22175, 22176, 22181, 22184, 22188, 22190, 22191, 22192, 22193, 22194, 22197, 22198, 22204 (全 61 力所)	従属集合数が多い 最寄り駅まで比較的近い 規模の小さい病院では参考としやすい病院となっている
22041	2 (1)	570 (24)	43 (158)	7 (65)	22089, 22102, 22124, 22143, 22144, 22181 (全 6 力所)	従属集合数が最小 医師数が最小 両入力項目の評価が高い
22055	7 (68)	160 (3)	97 (130)	8 (56)	22007, 22012, 22014, 22015, 22016, 22018, 22020, 22021, 22022, 22024, 22025, 22026, 22030, 22031, 22033, 22034, 22035, 22038, 22050, 22053, 22069, 22081, 22096, 22107, 22114, 22122, 22138, 22170, 22186, 22194, 22195, 22196, 22205 (全 33 力所)	最寄り駅まで近い 従属集合数が比較的多い
22061	18 (118)	180 (4)	99 (126)	14 (29)	22010, 22015, 22018, 22023, 22026, 22037, 22043, 22060, 22096, 22137, 22138, 22195, 22196, 22201 (全 14 力所)	最寄り駅まで近い
22097	3 (8)	1290 (58)	120 (108)	13 (33)	22009, 22051, 22052, 22092, 22102, 22104, 22106, 22125, 22143, 22148, 22177, 22181, 22190, 22193, 22197 (全 15 力所)	医師数が少ない
22101	5 (34)	270 (11)	220 (56)	2 (123)	22003, 22005, 22007, 22012, 22013, 22016, 22019, 22020, 22024, 22025, 22027, 22028, 22030, 22031, 22033, 22034, 22038, 22042, 22046, 22047, 22050, 22053, 22054, 22058, 22059, 22064, 22066, 22067, 22069, 22072, 22074, 22075, 22076, 22077, 22079, 22080, 22084, 22085, 22086, 22087, 22088, 22091, 22093, 22094, 22098, 22100, 22109, 22110, 22111, 22112, 22114, 22116, 22117, 22119, 22121, 22123, 22126, 22128, 22129, 22130, 22131, 22132, 22139, 22142, 22145, 22147, 22149, 22150, 22151, 22152, 22153, 22154, 22156, 22162, 22163, 22164, 22165, 22168, 22169, 22170, 22171, 22172, 22175, 22176, 22184, 22185, 22186, 22188, 22189, 22191, 22192, 22194, 22198, 22200, 22202, 22203, 22204, 22205, 22206 (全 99 力所)	従属集合が多い 規模の小さい病院では参考としやすい病院となっている 22036 と同様な傾向だが、出力項目の項目評価が反対となっている
22180	9 (84)	90 (1)	239 (52)	2 (123)	22007, 22010, 22015, 22018, 22020, 22023, 22024, 22026, 22030, 22031, 22034, 22037, 22043, 22053, 22069, 22096, 22137, 22138, 22186, 22195, 22196, 22201 (全 22 力所)	最寄り駅までが最も近い

表3 参照集合 (22029, 22036, 22101) を参照する DMU と比率

病院コード	22029	22036	22101
22003	0.2459	0.2451	1.3062
22019	0.1528	1.0354	0.4836
22027	0.7331	0.9523	0.0931
22028	0.6021	0.2261	0.1762
22042	0.0837	0.3813	0.2218
22046	0.211	1.0853	0.7598
22047	0.2995	0.1405	0.7422
22054	0.0536	0.2129	0.6658
22059	0.1066	0.3414	0.1753
22064	0.1421	1.007	0.193
22067	0.0495	0.1776	0.3436
22072	0.0001	0.3468	0.612
22076	0.0167	0.1647	0.191
22111	0.0087	0.3281	0.1093
22119	0.1383	0.8585	0.3213
22132	0.0404	0.3664	0.1707
22145	0.0404	0.1436	0.0619
22149	0.0084	1.2853	0.7837
22152	0.0786	1.4198	0.6137
22154	0.0792	0.1958	0.0043
22163	0.1637	0.0274	0.4173
22165	0.2044	0.0359	0.5167
22175	0.0375	0.129	0.6465
22176	0.1098	0.1876	0.2614
22184	0.1217	0.3504	0.0028
22188	0.0657	0.5343	0.272
22191	0.1639	0.219	0.1489
22192	0.0193	0.1146	0.3679
22198	0.0833	0.4999	0.2509
22204	0.2868	1.4491	1.1218

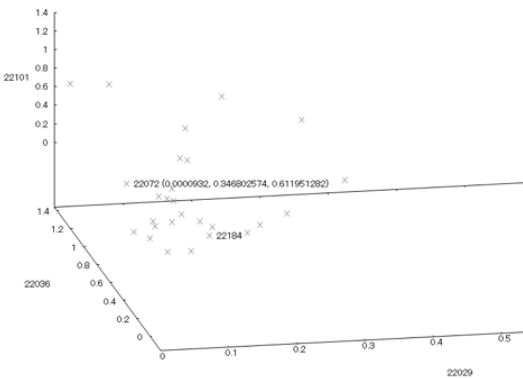


図8 表3の組み合わせを可視化したグラフ (3次元)

表4 参照集合 (22036,22055) を参照する DMU と比率

病院コード	22029	22101	病院コード	22029	22101
22005	0.357	0.642	22123	0.165	1.132
22013	0.318	1.012	22126	0.339	0.306
22058	0.378	0.403	22128	0.462	0.288
22066	0.09	0.241	22129	0.318	0.24
22074	0.329	0.348	22130	0.327	0.328
22075	0.172	0.642	22131	0.834	0.852
22077	0.209	0.293	22139	0.369	0.047
22079	0.09	0.927	22142	0.325	0.5
22080	0.435	0.074	22147	0.35	0.14
22084	0.508	0.335	22150	0.388	0.058
22085	0.336	0.189	22151	0.47	0.551
22086	0.427	0.042	22153	0.344	0.683
22087	0.229	0.038	22156	0.416	0.035
22088	0.399	0.838	22162	0.06	0.587
22091	0.349	0.275	22164	0.121	0.426
22093	0.276	0.325	22168	0.19	1.074
22094	0.091	0.034	22169	0.042	0.661
22098	0.409	0.002	22171	0.23	0.004
22100	0.191	0.335	22172	0.208	0.708
22109	0.336	0.248	22185	0.293	0.367
22110	0.21	1.099	22189	0.218	0.513
22112	0.296	0.22	22200	0.239	0.112
22116	0.257	0.188	22202	0.225	0.318
22117	0.429	0.402	22203	0.148	0.185
22121	0.07	0.377	22206	0.226	0.043

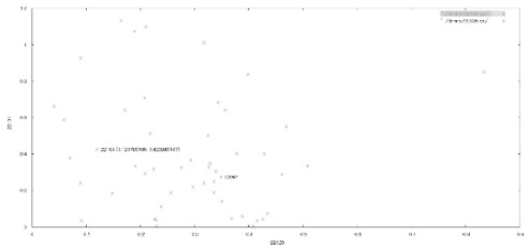


図9 表4の組み合わせを可視化したグラフ (2次元)

また静岡県内病院 158 件の同じデータを用いて k-means 法でのクラスタ化を実施した。ただしデータは単位に偏りがあるので偏差値に換算して実施している。また DEA で得られたクラスタ数が 9 であったため、k-means 法でもクラスタ数を 9 に指定している。表 5 は各クラスタに所属する DMU と、入出力項目毎の重心をまとめたものである。K-means 法では初期値に強く依存してしまい、クラスタが評価毎に変化する可能性もあり、また DEA で得られている改善指標のようなものがない。このことからビッグデータに対して DEA を用いて解析することで、多くの情報から解析が可能である。

表5 K-means 法でのクラスタ結果

医師数	最寄り 駅距離	病床数	診療科数	クラスタ	特徴
51.42	46.25	56.93	61.2	22032, 22044, 22046, 22053, 22149, 22152, 22201, 22204	中心地にある中規模病院グループ
46.43	47.5	52.87	43.1	22005, 22013, 22047, 22058, 22069, 22075, 22079, 22084, 22085, 22088, 22101, 22110, 22117, 22123, 22126, 22131, 22153, 22165, 22168, 22172, 22180, 22185, 22189	病床数が比較的多く、診療科数が少ない機能特化した病院グループ
67.15	49.08	68.96	70	22007, 22014, 22016, 22020, 22023, 22024, 22025, 22026, 22030, 22031, 22034, 22037, 22043, 22196	医師数、ベッド数、診療科数が良い中・大規模病院グループ
46.4	81.69	47.35	49.42	22009, 22011, 22027, 22028, 22106, 22120, 22157, 22158, 22174	最寄り駅が非常に遠い他の項目は数値が低い通院が難しい病院グループ
96.72	48.64	81.1	72.01	22010, 22015, 22137, 22138	医師数、病床数、診療科数が突出している。大病院グループ
46.32	44.06	42.74	43.89	22022, 22041, 22042, 22052, 22054, 22066, 22067, 22072, 22076, 22077, 22081, 22087, 22100, 22104, 22111, 22121, 22122, 22132, 22145, 22154, 22162, 22164, 22169, 22175, 22176, 22184, 22186, 22188, 22190, 22191, 22192, 22193, 22194, 22198, 22202, 22203, 22205	すべての項目に対して数値が低い中心地にある小病院グループ
47.25	43.42	44.49	53.72	22033, 22035, 22036, 22038, 22050, 22055, 22060, 22061, 22064, 22089, 22092, 22096, 22097, 22102, 22107, 22114, 22119, 22124, 22125, 22143, 22144, 22148, 22170, 22181, 22195	診療科数が比較的多い他の項目は数値が低い中心地にある通院向け病院
46.01	53.88	45.09	43.73	22051, 22059, 22068, 22074, 22086, 22091, 22109, 22112, 22116, 22129, 22130, 22142, 22149, 22166, 22171, 22177, 22197, 22200, 22206	最寄り駅が遠い他の項目は数値が低い中心地から離れた小病院
46.01	61.81	47.75	41.64	22080, 22093, 22094, 22095, 22098, 22108, 22128, 22139, 22147, 22150, 22151, 22156, 22199	最寄り駅が非常に遠い他の項目は数値が低い遠隔地にある小病院

3.3 類型化解析基盤における計算時間

今回提案している類型化解析基盤では、ビッグデータの環境に対し大規模な類型化が実施可能である。提案した解析基盤の計算時間に対する基準を示すため、ここでは DMU 数に対してどの程度計算時間が必要となるか算出する。ビッグデータを扱うとき得られたデータを効率的にデータベースに格納し、可視化につなげることが重要であることから、計算だけでなくデータベース格納時間まで含め計測している。表 6 は CPU 最大コア数:32, CPU 周波数:4000MHz, メモリ:32G バイト, HDD:300G バイトのサーバで計算した結果である。この結果より、並列数を増やしても入出力のオーバーヘッド等で計算速度効果は薄れてしまうことがわかる。並列数を 1 サーバ内での並列計算数は 4 程度が適当である。図 10 は今回のような DEA を計算する

ときに、理想的なサーバ構成となる。すべてのクラスタで同じ DMU 情報を共有し、各クラスタに計算を振り分ける。並列数増加での効果が頭打ちであることから、このようなクラスタ構成が望まれる。

この時新たに生成されるレコードは、図 3 の DMU 数を n 、入力項目数を m 、出力項目数を s とすると、計算テーブルは 1 レコード、主問題結果テーブルは n レコード、ウエイトテーブルは $n(m+s)$ レコード、双対問題結果テーブルは n レコード、双対問題入値テーブルは n^2 レコードとなる。よって主問題でのレコード数は $n^2(m+s)$ レコード、双対問題でのレコード数は n^3 レコードとなる。双対問題入値テーブルの挿入方法を multiple insertion[24]などで工夫することで、計算時間の短縮を図ることができる。

DEA計算サーバ構成

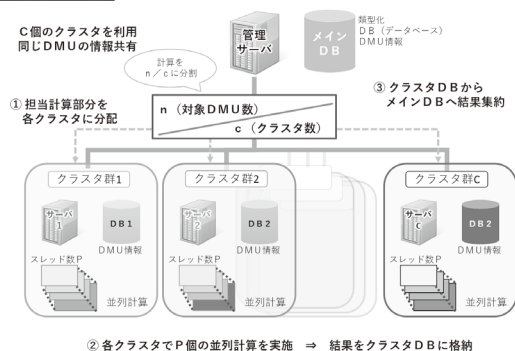


図 10 DEA 計算サーバ構成

表 6 並列計算数を変えた場合の計算時間 (m=s=2)

DMU 数\並列数	1	4	8	16
5000	4208	1626	1598	1589
10000	36534	14661	13106	12739

またクラウド・コンピューティングの環境を利用し、柔軟にスケールアップやスケールアウト [25] を実施することで、容易に計算時間の短縮を図ることが可能である。全国の病院で類型化を行なった場合 7,000 程度の DMU 数となるが、これを 2 種類のサーバを用意して計算時間の比較を実施した。使用したサーバの環境は Amazon Web Service[26] の EC2 である。表 7 は今回利用した 2 サーバのスペックである。サーバの OS は両方とも CentOS6.5 64bit 版を用いている。

表 7 利用サーバのスペック比較

サーバ種類	OS	アーキテクチャ	vCPU	メモリ (GiB)
m1.small	CentOS6.5	64bit	1	1.7
m1.xlarge	CentOS6.5	64bit	4	15

m1.small での DMU7713 件での類型化計算時間は 7906 秒に対し、m1.xlarge での計算時間は 4233 秒であり 90 分を下回る計算時間で計算可能である。図 11 は類型化計算時間の比較を示している。DMU 件数が増えても、サーバスペックを上げることで相当数の DMU まで計算可能である。

4 さ い ご に

本研究では、ビッグデータ環境に対応可能な汎用的な類型化解析基盤として DEA を用いたものを提案し、データベース及びシステム構成を示した。またその結果を可視化する手法やこの類型化解析基盤を用いた場合の計算時間基準を示し、実用的に利用可能なシステ

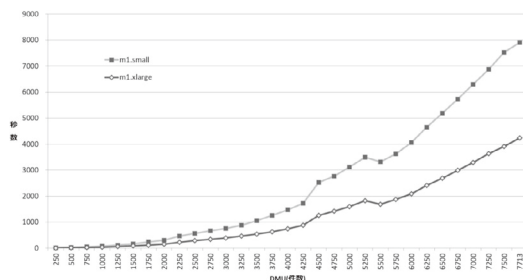


図 11 類型化計算時間の比較

ムとして提案した。現在多方面で利用されているクラウド・コンピューティング環境を利用することで、大規模な計算にも対応可能である。このような枠組みで計算することで、多角度的の評価を大規模に実施することができる。また病院類型化を実在するデータを用いて日本全国規模で実施し、参照集合を算出してどのように分類されるかを示した。本研究の類型化解析基盤を用いることで、多岐にわたるモデルを標準化された枠組みの中で類型化して代表的な DMU を明らかにし、今後の方向性を見出せると期待する。

謝辞

本研究は、平成 27 年度 厚生労働科学研究費補助金地域医療基盤開発推進研究 事業「大規模ネットワーク AHP を用いた「医療の質」に関わる情報が患者の医療 機関選択に与える影響に関する検証」（研究代表者：藤澤由和）による、研究成果の一部を取りまとめたものである。本論文を作成するにあたり、静岡産業技術専門学校篠原美樹さんは全体モデル作成と類型化計算例において、静岡理科大学水野研究室の澤木みゆさんは類型化計算の集計において、静岡大学の関睦実さんは論文作成において、それぞれ大変ご尽力いただきました。ここに感謝の意を表します。

参 考 文 献

- [1] 山端 博：“ビジネス・インテリジェンスと CRM：データマイニング・ビジネスの実際”，オペレーションズ・リサーチ：経営の科学, Vol. 43, No. 12, pp. 653-657 (1998)
- [2] 小出 治, 大坂谷吉行：“クラスター分析による行政投資パターンからみた都道府県の類型化”，地域学研究：日本地域学会年報, No. 7, pp. 113-137 (1976)
- [3] 廣井 悠：“福島原子力発電所からの避難行動に関する調査と分析”，都市計画論文集, Vol. 49, No. 3, pp. 537-542 (2014)
- [4] 野村総合研究所 IT 事業推進部：“ジオデモグラフィック分析手法によるアプローチ”，知的資産想像 (2008 年 11 月号), p. 86 (2008)

- [5] 伊藤武彦, 中山いづみ: “小学校通学区 (学区) の類型化の試み — 姫路市をモデルにして —”, 岡山大学大学院教育学研究科研究集録, No. 158, pp. 23-26 (2015)
- [6] 安達一寿: “ブレンディッドラーニングでの学習活動の類型化に関する分析”, 日本教育工学会論文誌, Vol. 31, No. 1, pp. 29-40 (2007)
- [7] 因幡雄起, 大石博之, 杉原成満, 竹本大昭, 古川浩平: “DEA の大規模問題適用システムの構築とそれを用いた土砂災害危険箇所順位の設定”, 砂防学会誌, Vol. 59, No. 1, pp. 21-28 (2006)
- [8] 末吉俊幸: 「DEA—経営効率分析法—」, 朝倉書店 (2001)
- [9] 井階美歩, 逸見宣博, 中川慶一郎, 生田目崇, 田辺隆人: “NUOPT を用いた大規模 DEA プログラムの開発”, 日本オペレーションズ・リサーチ学会秋季研究発表会アブストラクト集, pp. 320-321 (2003)
- [10] 伊藤直也, 勝見祐己, 田中慎司, ひろせまさあき, 安井真伸, 横川和哉: 「サーバ/インフラを支える技術」, 技術評論社 (2009)
- [11] lp_solve reference guide menu, <http://lpsolve.sourceforge.net/5.5/>, 2015/12
- [12] 関谷和之: “A Two-stage Approach for Russell Measure in DEA”, 数理解析研究所講究録, No. 1629, pp. 152-163 (2009)
- [13] Charnes, A., Cooper, W.W., Lewin, A.Y. and Seiford, L.M.(Eds.), 刀根 薫, 上田 徹 (監訳): 「経営効率評価ハンドブック—包絡分析法の理論と応用—」, 朝倉書店 (2007)
- [14] Sugiyama, M. and Yamada, Y.: “Data Envelopment Analysis Using Virtual DMU as Intermediates: An Application to Business Analysis of Japan’s Automobile Manufacturers”, *J. Jpn. Ind. Manage. Assoc.*, Vol. 50, No. 6, pp. 341-354 (2000)
- [15] Charnes, A., Cooper, W. W. and Rhodes, E.: “Measuring Efficiency of Decision Making Units”, *Eur. J. Oper. Res.*, Vol. 2, pp. 429-444 (1978)
- [16] 刀根 薫: 「経営効率性の測定と改善—包絡分析法 DEA による」, 日科技連 (1993)
- [17] 新村秀一: “DEA による回帰型データのクラスター分析”, 成蹊大学一般研究報告, 第 45 巻, pp. 1-37 (2011)
- [18] 山内秀敏, 陳 湘玉, 池上敦子, 八巻直一: “DEA を用いたレーダーチャートの特徴分類”, 経営情報学会東海支部大会, <http://hdl.handle.net/10297/6700>(2012)
- [19] Cook, W.D. and Zhu, J., 森田 浩 (訳): 「データ包絡分析法 DEA」, 静岡学術出版 (2014)
- [20] 杉山 学: 「経営効率分析のための DEA と Inverted DEA」, 静岡学術出版 (2010)
- [21] HeartRails Geo API, <http://geoapi.heartrails.com/api.html#nearest>, 株式会社ハートレイルズ (2015/12/01 参照)
- [22] gnuplot homepage: <http://www.gnuplot.info/>, (2016/08/01 参照)
- [23] The Google Maps Geocoding API, <https://developers.google.com/maps/documentation/geocoding/intro>, Google Developers (2015/12/01 参照)
- [24] Stack OverFlow: <http://stackoverflow.com/> (2016/08/20 参照)
- [25] 佐々木拓郎, 林晋一郎, 小西秀和, 佐藤 瞬: 「Amazon Web Services」, SB クリエイティブ (2015)
- [26] Amazon Web Service, <https://aws.amazon.com/jp/>, Amazon Web Services, Inc (2015/12/01 参照)