

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

総合型類型化解析基盤の 構築とその応用

**Construction of a Comprehensive Analysis Platform
for Typology and its Application**

October 15, 2024

1 本研究の背景

2/22

背景

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、労働者不足など社会は大きな問題を抱えている。こ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ごとの特徴を明らかにし、効率的な改善を提示することが必要である。

目的

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1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を用いられていたが、特徴を明示しやすい DEA(Data Envelopment Analysis) を用いる。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を検証していく。

- 類型化：共通の性質・特徴をもつものどうしをまとめた一つの型。
- 特徴抽出：元の観測データの特徴ベクトルから、変換処理や選択処理を経て予測に有用な特徴をつくったり、可視化されてデバッグしやすい特徴を抽出すること。

2.1 類型化解析フロー

3/22

類型化を実施するには次のようなフローになる。

- 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、線型計画法を計算可能な環境を設定
- 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、参照集合を算出してクラスタ毎分類
- ⑥ 類型化の表示
 - API：ソフトウェアやプログラム同士をつないでいるインターフェース
 - 参照集合：見本とすべき DMU の集合

1 はじめに

2 類型化解析基盤
の構成

3 数値計算例

2.1 類型化解析フロー

4/22

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

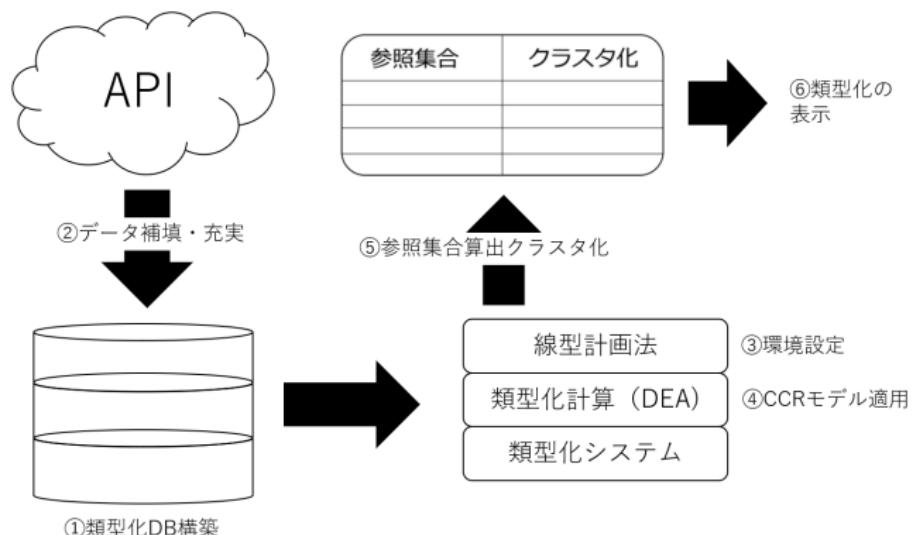


図 1: 類型化解析のフロー

2.2 類型化データベースの作成と API の利用

5/22

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

類型化データベースは基本的に図 2 のような形で作成する。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。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

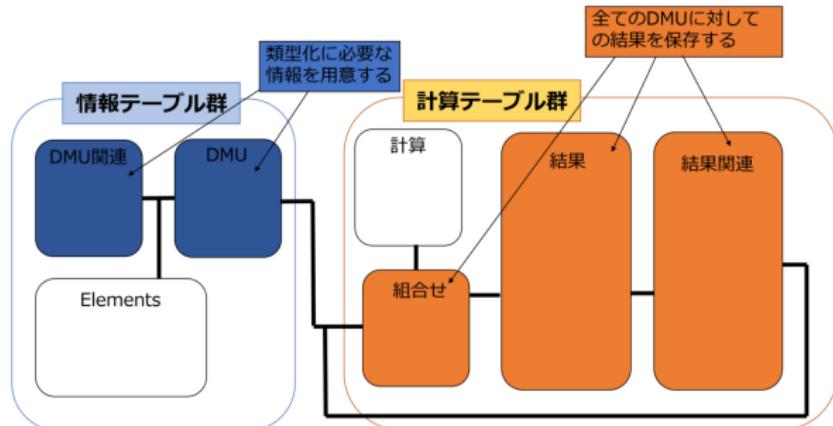


図 2: 類型化解析のフロー・類型化データベース構造

2.2 類型化データベースの作成と API の利用

6/22

- 1はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

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 3 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作成し API から情報を取得できる関数を呼び出す。

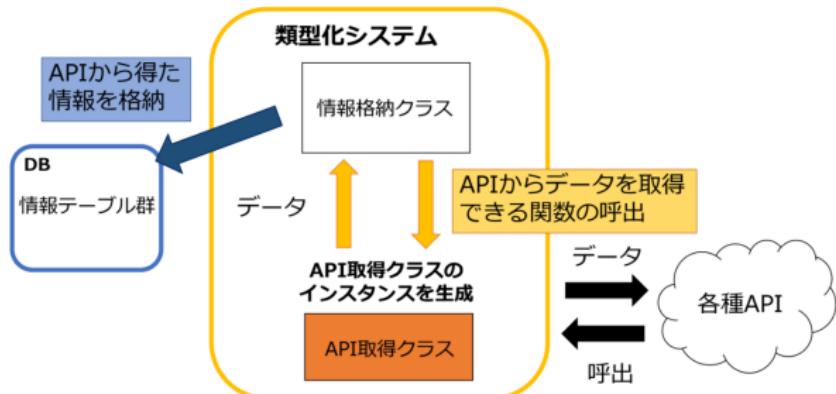


図 3: API から情報を取得するフロー

2.3 DEA の利用

7/22

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

今回は DEA 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とらえることのできる CCR モデルを利用していく。

DEA とは

効率性を分析する方法の 1 つであり、公共機関や民間企業などを評価するために利用されている。

- 複数項目での総合評価
- 個性的で多様性を活かした評価
- 改善値の定量的な把握

2.3 DEA の CCR モデル定式化

8/22

CCR モデルとは

- 最初の DEA のモデル
- 効率値が 1 以下の制約下で DMU にとって最も大きくなるようにする比率形式モデルをもとにする
- それを線形計画法で記述するモデル

- DMU...DEA で効率性を評価する際に、対象とする評価単位 (組織)
- 比率形式モデル... 仮想的出力と仮想的入力の比を使って比較するモデル
- 線形計画法..... 制約条件のもとで目的関数を最大・最小にする値を求める手法

2.3 DEA の CCR モデル定式化

9/22

- 1はじめに
- 2類型化解析基盤の構成
- 3数値計算例

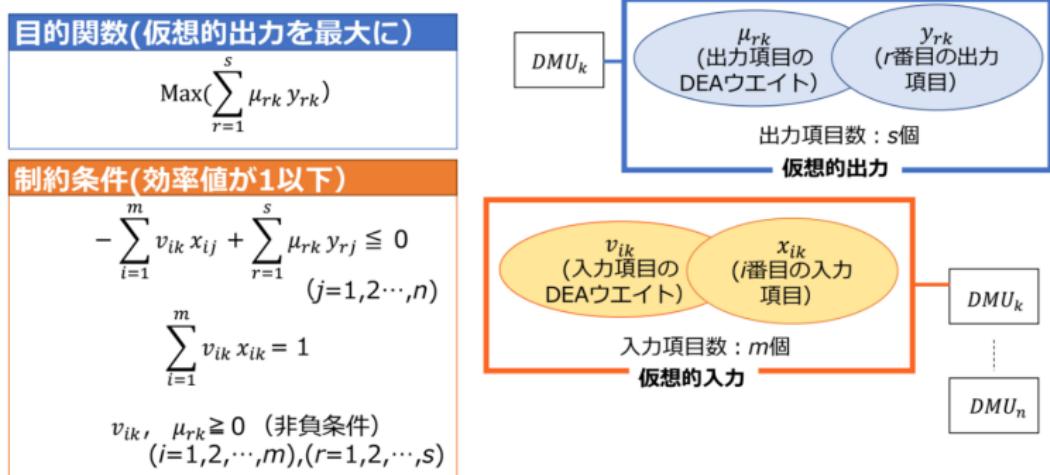


図 4: $DMU_k (k = 1, 2 \dots n)$ に対して DEA 効率値を計算する場合の主問題

2.3 DEA の CCR モデル定式化

10/22

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

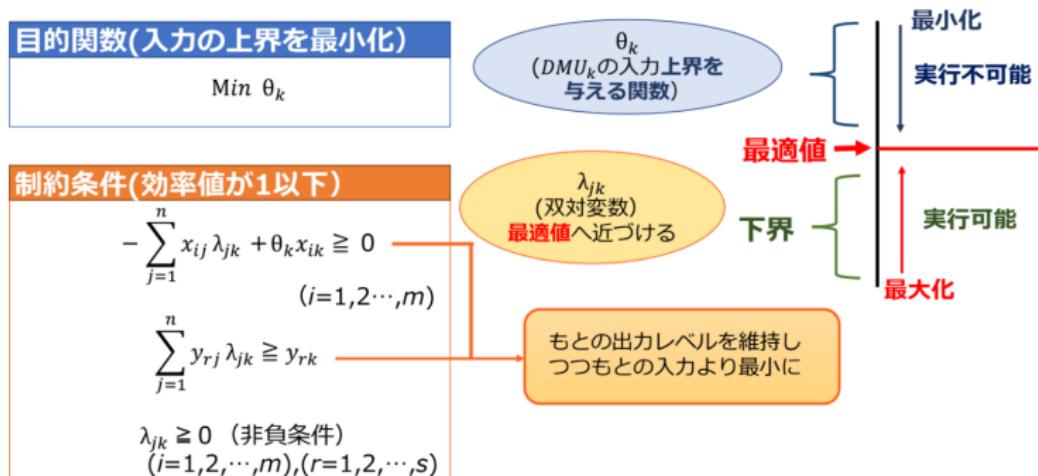


図 5: $DMU_k (k = 1, 2 \dots n)$ に対して DEA 効率値を計算する場合の双対問題

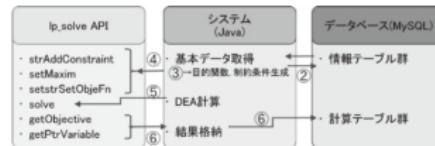
2.3 Ip_solve の利用

11/22

- 1はじめに
- 2類型化解析基盤の構成
- 3数値計算例

DEA を CCR モデルで扱うことで、線型計画法として計算を実施する。今回は Ip_solve を採用した。

- 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



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

- ① プログラミング言語、データベースの設定
- ② システムから情報テーブル群にアクセスしデータを取得
- ③ 取得データから目的関数、制約条件を作成
- ④ Ip_solveに目的関数、制約関数を設置
→目的関数をすべてのDMUの条件で実施
(制約条件は同じものを利用)

図 6: DEA の全体計算を行う過程

2.3 参照集合の算出と DMU の分類

12/22

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

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。今回は明確なプロセスで表現される双対問題のウエイトを用いた方法で DMU の分類を行う。

- ① 主問題及び双対問題の目的関数を計算し目的関数=1 となる DMU を算出
→ 参照集合を形成し、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 $\lambda_{jk} = 0$ となるとき
→ 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

3.1 静岡県内病院での計算例

13/22

静岡県内病院で入力項目・出力項目すべてを取得できた病院 158 件を対象に DEA 計算を行う。

1 はじめに

2 類型化解析基盤の構成

3 数値計算例

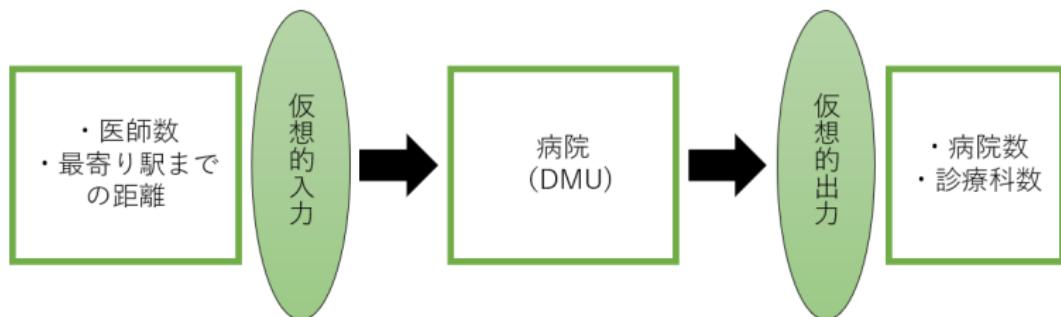


図 7: 入力項目と出力項目

3.1 静岡県内病院での計算例

14/22

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

参照集合	医師数	最寄駅距離	病床数	診療科数	クラスタ	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, 22178, 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)	最寄駅まで近い 両出力項目の評価は比較的高い	
22036	5 --	330 --	99 --	8 --	22003, 22009, 22012, 22014, 22016, 22019, 22021, 22022,	従属集合数が多い	

図 8: DEA 計算結果から類型化

- 第 2 列から第 5 列：上段（用いたデータ）、下段（全 DMU 内の順位）
- 第 2 列から第 5 列：入力項目と出力項目であり、値を降順に並べた順位
- 22029：医師数最小であり、出力 2 項目は比較的いい数値となっている。参照している病院数が 102 と最大となっている。

3.1 静岡県内病院での計算例

15/22

- 1はじめに
- 2類型化解析基盤の構成
- 3数値計算例

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	クラスタ	特徴	
22029	2	3179	400	18	22003, 22005, 22009, 22011, 22013, 22019, 22027, 22028, 使医集合数が最大数		
(1)	(112)	(22)	(21)		22042, 22046, 22047, 22051, 22052, 22054, 22058, 22059, 医師数が最小		
					22064, 22066, 22067, 22068, 22072, 22074, 22075, 22076, 出力項目の2項目の評価		
					22077, 22079, 22080, 22084, 22085, 22086, 22087, 22088, は高い、		
					22091, 22092, 22093, 22094, 22095, 22096, 22100, 22104,		
					22106, 22108, 22109, 22110, 22111, 22112, 22116, 22117,		
					22119, 22120, 22121, 22122, 22125, 22126, 22128, 22129,		
					22130, 22131, 22132, 22136, 22142, 22145, 22147, 22148,		
					22149, 22150, 22151, 22152, 22153, 22154, 22156, 22157,		
					22158, 22162, 22163, 22164, 22165, 22166, 22168, 22169,		
					22171, 22172, 22174, 22176, 22176, 22177, 22184, 22185,		
					22188, 22189, 22190, 22191, 22192, 22193, 22197, 22198,		
					22199, 22200, 22202, 22203, 22204, 22206, (全 102 方所)		
22032	50	140	350	13	22010, 22023, 22037, 22040, 22044, 22060, 22137, 22201, 最寄り駅まで近い		
(34)	(137)	(2)	(25)	(33)	(全 8 方所)		
					両出力項目の評価は比較的高い		
22036	5	330	99	8	22003, 22009, 22012, 22014, 22016, 22019, 22021, 22022, 使医集合数が多い		
(34)	(12)	(126)	(56)		22025, 22027, 22028, 22033, 22035, 22038, 22042, 22046, 最寄り駅まで比較的近い		
					22047, 22050, 22051, 22052, 22054, 22059, 22064, 22067, 病種の小さい病院では参考		
					22072, 22076, 22081, 22089, 22092, 22102, 22107, 22111, 考えしやすい病院となつている		
					22114, 22119, 22122, 22124, 22125, 22132, 22143, 22144,		
					22145, 22148, 22149, 22152, 22154, 22163, 22166, 22170,		
					22175, 22176, 22181, 22184, 22188, 22190, 22191, 22192,		
					22193, 22194, 22197, 22198, 22204, (全 61 方所)		
22041	2	570	43	7	22089, 22102, 22124, 22143, 22144, 22181, 使医集合数が最小		
(1)	(24)	(158)	(65)		医師数が最小		
					両出力項目の評価が高い		

図 9: 類型化参照集合の一部

それぞれ何らかの特徴を持つ参照集合となっている

同時に、それらを参照する從属集合が分かる

3.2 類型化解析結果と可視化

16/22

病院類型化の結果から、ある日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ものである。

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

図 10: 参照集合を参照する DMU と比率

1はじめに

2 類型化解析基盤
の構成

3 数値計算例

3.2 類型化解析結果と可視化

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 を利用した。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。

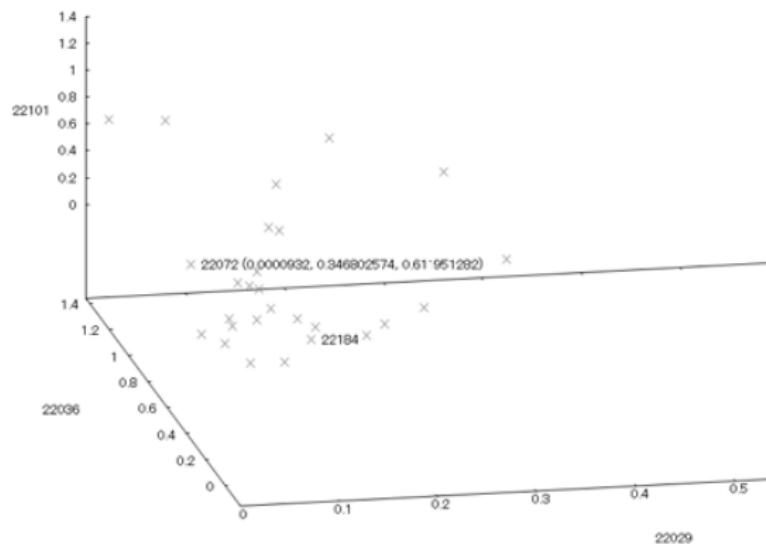


図 11: 類型化解析結果の可視化

3.2 k-means 法でのクラスタ化

18/22

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。

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.	医師数、ベッド数、診療数が良い

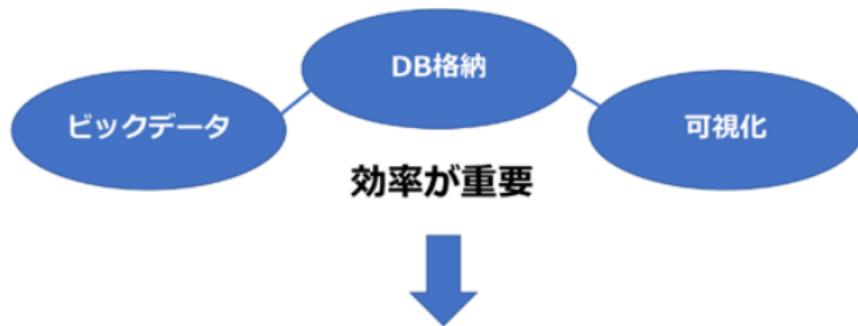
図 12: k-means 法でのクラスタ結果

提案手法と比べて分かったこと

- K-means 法は初期値に強く依存してしまう
- クラスタが評価ごとに変化する可能性がある
- DEA で得られている改善指標のようなものがない。

3.3 類型化解析基盤における計算時間

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例



提案した解析基盤の計算時間に対する基準を示す
ここでは、DMU数に対してどの程度の計算時間が必要となるかを算出

図 13: 類型化解析における計算時間

3.3 類型化解析基盤における計算時間

20/22

DEA計算サーバ構成

- 1 はじめに
- 2 類型化解析基盤の構成
- 3 数値計算例

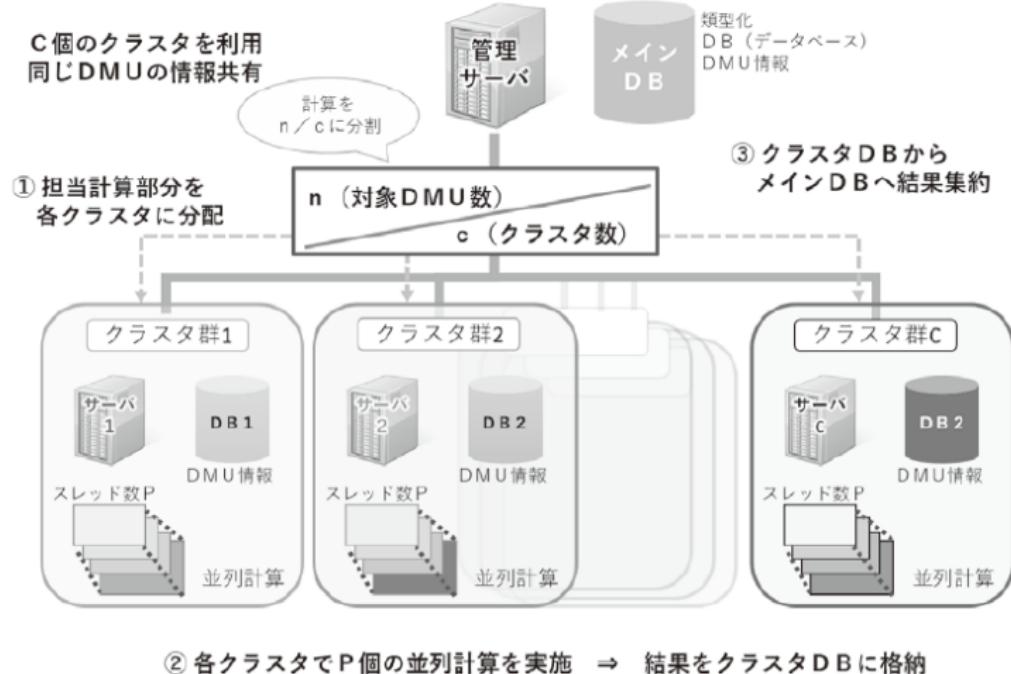


図 14: DEA の計算サーバ構成

3.3 類型化解析基盤における計算時間

21/22

全国の病院で類型化を行った場合 7,000 程度の DMU 数となるが、2 種類のサーバを用意して計算時間の比較を実施した。

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

図 15: 利用サーバのスペック比較

- m1.small での 7713 件での類型化計算時間は 7906 秒
- m1.xlarge での計算時間は 4233 秒

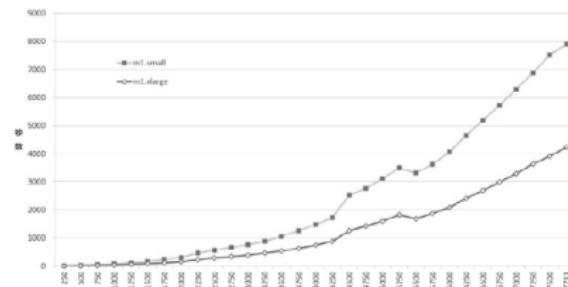


図 16: 類型化計算時間の比較

1 はじめに

2 類型化解析基盤
の構成

3 数値計算例

4 おわりに

本研究では

ビックデータ環境に対応可能で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た

- DEA の利用：相対的な順位だけでなく各項目の具体的な改善値も把握できる
- 結果の可視化：改善のためのディスカッションに役立てることができる
- 計算：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

様々な分野の代表的な DMU を明らかにし、今後の方向性を見出せると期待する。