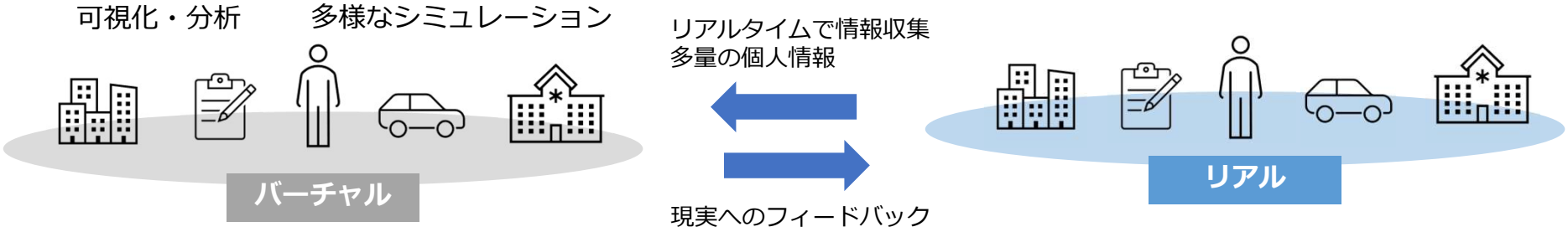
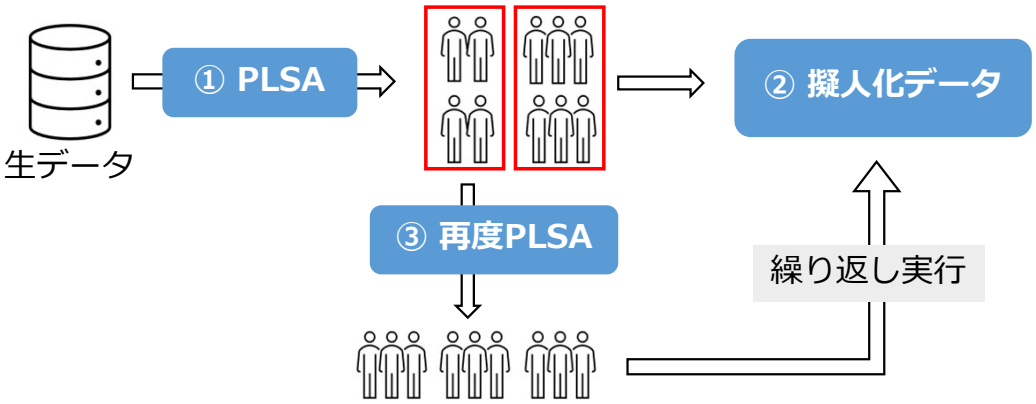


確率的潜在意味解析による プライバシー保護情報作成方法の提案

管原 侑希[1,2]，櫻井 瑛一[1]，本村 陽一[1]，信夫 咲希[2]，岡田 幸彦[3,4]，塚尾 晶子[5]，久野 譜也[6]

1 国立研究開発法人 産業技術総合研究所 人工知能研究センター
2 筑波大学大学院 サービス工学学位プログラム
3 筑波大学 システム情報系
4 筑波大学 人工知能科学センター
5 株式会社つくばウエルネスリサーチ
6 筑波大学 スマートウエルネスシティ政策開発研究センター

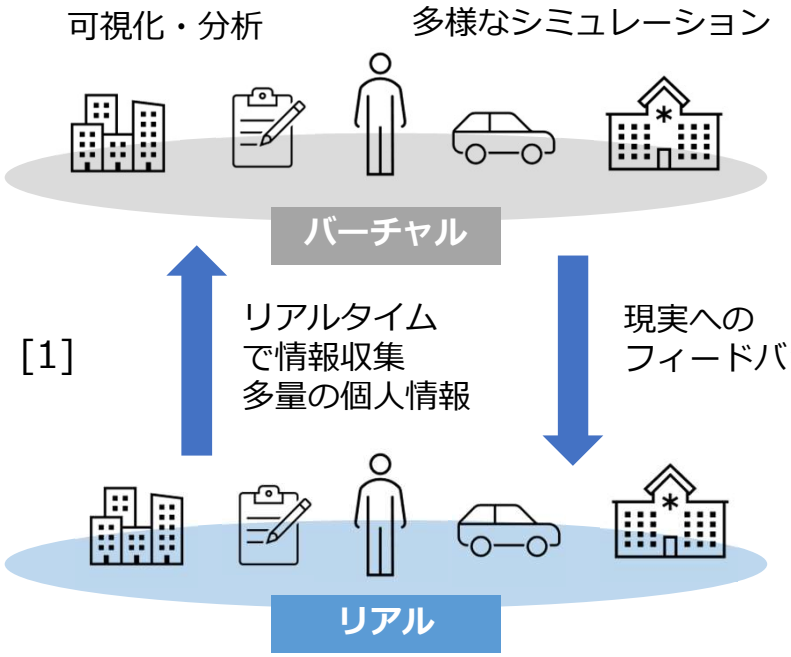


研究背景

デジタルツイン

…データやAIを活用して新たな価値を
創出するための手段

- 多様な生活者の行動や意識の変化などへの発展 [1]
- 個人のライフスタイルなどをリアルタイムに反映する**生活者デジタルツイン**の構築
  →ライフスタイルや行動パターンが個人の健康状態に与える影響が観察可能に[2]



生活者デジタルツイン

生活者デジタルツインによるサービス提供には**個人情報の保護が重要**[3]
さらにサービス設計においては**人々の行動の異質性**も考慮する必要[4]

クラスタリングによる解析

ミクロアグリゲーション

…ミクロデータを k 個のレコードを有する同質的なレコード群にグループ化した上で、
そのレコードにおける個々の属性値を平均値等の代表値に置き換える匿名化手法[5]

ミクロアグリゲーションを用いた研究

著者	対象データ	目的
伊藤ら（2014）[6]	全国消費実態調査と家計調査の個票データ	ミクロデータに対する匿名化技法の適用可能性の検証
井手ら（2017）[7]	日本老年学的評価研究 (JAGES) による 10 万人以上の高齢者アンケートデータ	各個人の郵便番号レコードによるミクロアグリゲーション を行い、確率的潜在意味解析を用いた匿名化情報の分析

人々の異質性の考慮という観点から、**クラスタリングによる解析を想定した**うえで、ミクロアグリゲーションを用いて行われた実証研究は未だ少ない

研究目的

生活者デジタルツインの構築に向け、確率的潜在意味解析（PLSA）を用いた新たなプライバシー保護情報の作成方法を提案すること

生活者デジタルツインの構築に向けては、**個人情報の保護**・
人々の異質性を考慮したサービス提供が必要

人々の異質性を考慮するには、**クラスタリングによる解析**が有効

➡ 確率的潜在意味解析（PLSA）を用いたクラスタリング

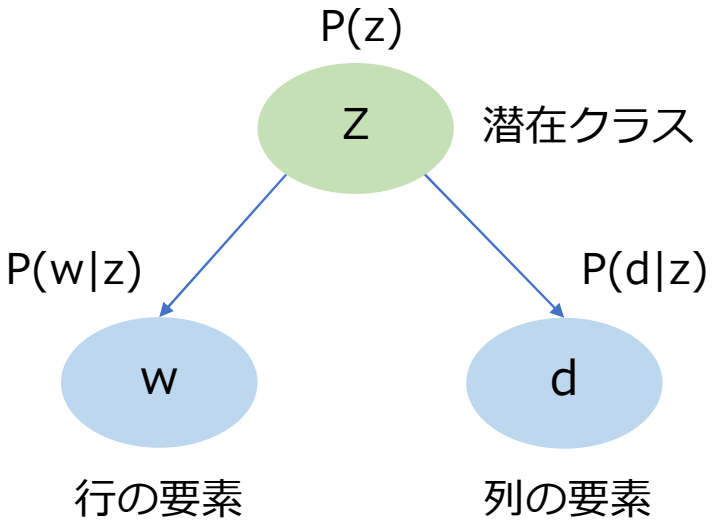
⬇

個人情報を保護したうえでも、PLSAによるクラスタリングを有効に行える必要

確率的潜在意味解析（PLSA）とは

- 文書分類の手法としてHofmann[9]により提唱
  文章d中の単語wが潜在変数zを介して生成されると仮定し、
  EMアルゴリズムによる尤度最大化で、共起するデータに
  潜在変数zを付随

- アンケートの回答者wが質問項目dに潜在変数zを介して
  回答すると仮定して、**アンケートの回答が似た人を類型化することに応用**



他のクラスタリング手法と比べて、データの共起関係を考慮し、
似た人々を類型化することが可能

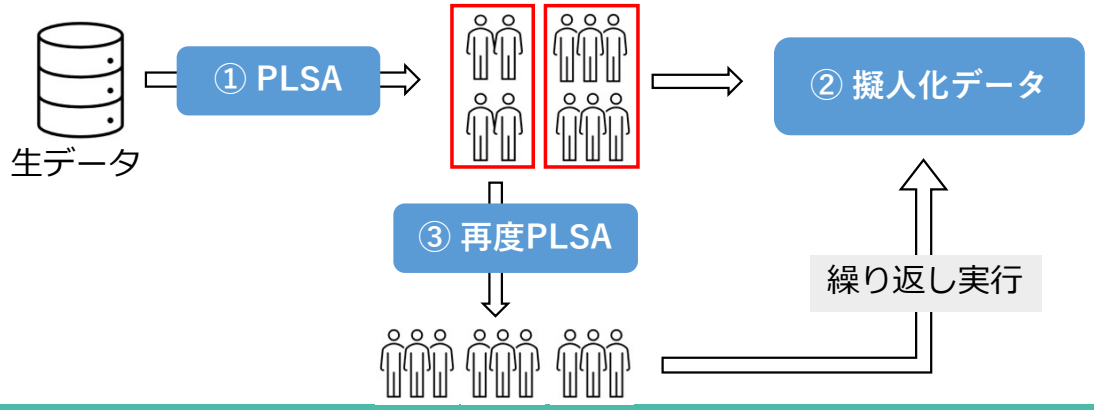
提案手法

① 生データに対して、十分に大きなクラス数のもとでPLSAを実行し、生データを複数のクラスタに類型化

② 所属人数が3人以上20人以下となったクラスタを**1件の擬人化データ**として採用

③ それ以外のデータを用いて再度PLSAを実行

④ 2,3のプロセスを全てのクラスタの所属人数が**3人以上20人以下**になるまで繰り返し行う



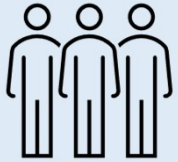
例）Z001に所属する人が4人であった場合

ID	アンケート項目				所属クラスタ
	①	②	③	…	Z001
A	1	0	0	…	1
B	0	1	0	…	1
C	1	0	0	…	1
D	0	0	1	…	1

回答を合計

ID	アンケート項目				所属クラスタ
	①	②	③	…	Z001
擬人A	2	1	1	…	ー

使用データ

	自治体Aのアンケートデータとその回答者の介護保険・レセプトデータを統合したデータ
	自治体Aにおける45歳以上の国民健康保険および後期高齢者医療保険の加入者
	生活状況・身体状況・性格・健康関心度・健康状態・居住地域学歴・就労状況・経済状況などに関する計46設問

データ数

	使用データ数(件)
収集データ	4,045
擬人化データ作成 *1	2,270
フレイル該当率の比較 *2	1,953

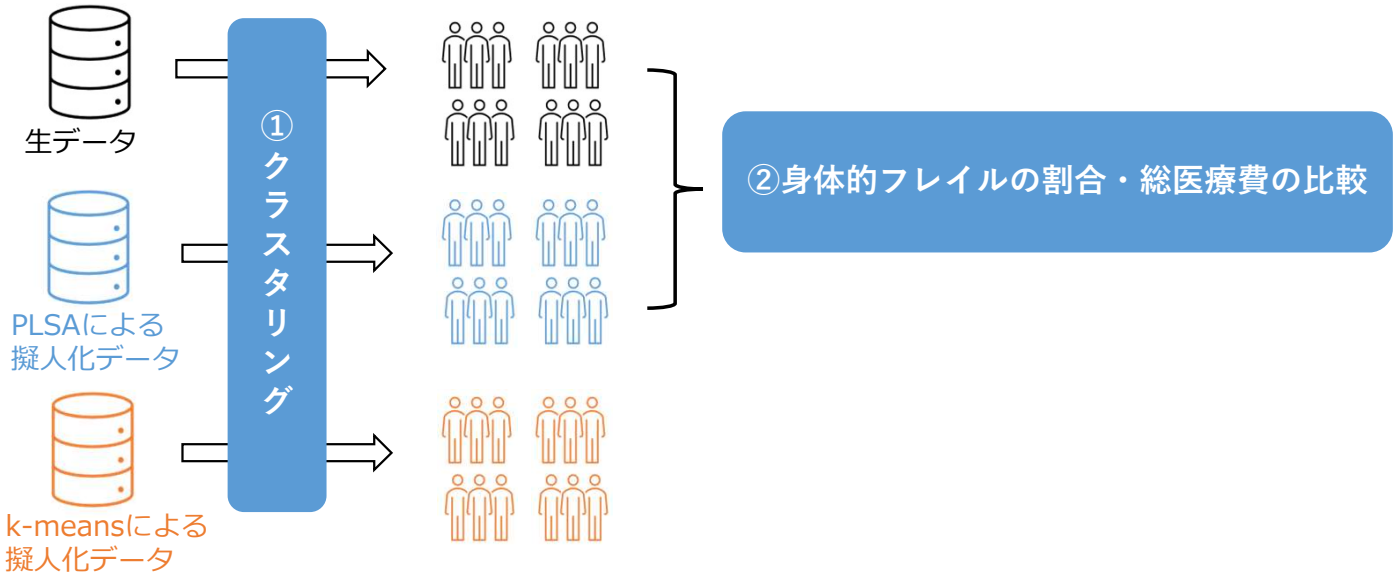
*1 社会的孤立に関するアンケート項目に全て回答し、数値回答項目を除くその他の項目に欠損数10以内で回答

*2 身体的フレイルの判定に必要な項目に回答し、2011~2019年の間に介護判定を受けていない

提案手法の評価方法

全アンケート項目を用いて提案手法およびk-means法の擬人化データを作成
① 生データ・提案手法による擬人化データ・k-means法による擬人化データのそれぞれで、
社会的孤立に関する設問でクラスタリングを行い、特徴を比較[8]
② 生データおよび提案手法による擬人化データから得られたセグメントごとに、
身体的フレイルの割合や総医療費を比較

イメージ図



擬人化データの作成結果

計24回のPLSAの試行により、2,270件の生データから**299件の擬人化データ**を生成

PLSA試行回数	使用データ											得られたセグメント数
PLSA1回目	2,270件										⇒	15セグメント
PLSA2回目	2,167件										⇒	18セグメント
PLSA3回目	2,043件										⇒	26セグメント
PLSA4回目	1,778件										⇒	13セグメント
PLSA5回目	1,703件										⇒	8セグメント
…	…											…
PLSA20回目	371件										⇒	11セグメント
PLSA21回目	308件										⇒	12セグメント
PLSA22回目	240件										⇒	18セグメント
PLSA23回目	128件										⇒	25セグメント
PLSA24回目	28件										⇒	1セグメント
計												**299セグメント**

結果

PLSAによる社会的孤立のクラスタリング結果

※各クラスタには幸福度の高低で名称を付与

- 生データ、提案手法による擬人化データ、k-means法による擬人化データの全てのデータセットにおいて**4クラス**が採用（AIC基準より）
- 提案手法による擬人化データクラスタでは、**生データと似た特徴が抽出された**一方、k-means法による擬人化データでは、生データと似た特徴を抽出できず

生データの場合

クラスタ	各クラスタに所属する代表的なアンケート項目
S1	社会活動参加あり 地域貢献意識_高い
S2	家族以外との会話頻度_普通でない，独居でない， 家族以外との会話頻度_高い， 家族との会話_毎日，誰かと食事あり
S3	電話やSNSでの会話頻度_高くない，独居でない， 家族以外との会話頻度_普通，誰かと食事あり 家族以外との会話頻度_高くない
S4	地域貢献意識_低い

提案手法を用いた場合

クラスタ	各クラスタに所属する代表的なアンケート項目
C1	社会活動参加あり 地域貢献意識_高い
C2	家族以外との会話頻度_高い，誰かと食事あり， 家族との会話_毎日，地域協力信頼_普通 婚姻状態_配偶者あり
C3	地域協力信頼_普通，誰かと食事あり， 家族との会話_毎日，婚姻状態_配偶者あり， 電話やSNSでの会話頻度_普通
C4	地域貢献意識_低い

k-means法を用いた場合

クラスタ	各クラスタに所属する代表的なアンケート項目
W1	社会活動参加あり
W2	家族以外との会話頻度_高い，地域行事の参加， 地域貢献意識_高い，誰かと食事あり， 趣味関係の活動参加
W3	地域行事の参加，社会活動_その他参加， 地域貢献意識_高い，その他のボランティア活動の参加， 学習・教養サークルの参加
W4	地域貢献意識_高い

セグメントごとの総医療費の比較

生データから得られたクラスタ間の違いと
同様の傾向を**総医療費の平均値においては**確認することができた

- 総医療費：2017年から2019年の3年間の平均医療費を算出

生データの場合

クラスタ	N数（0円の人を除く）	総医療費（千円）	
		平均値	標準偏差
S1	167	27.9[a]	30.7
S2	767	30.0[a]	48.8
S3	948	34.5[a]	53.8
S4	313	39.3[b]	53.8
計	2,195	33.1	50.8

提案手法を用いた場合

クラスタ	N数（0円の人を除く）	総医療費（千円）	
		平均値	標準偏差
C1	51	26.3[a]	13.6
C2	80	30.0[a]	16.2
C3	141	31.7[a]	18.7
C4	27	42.7[b]	23.7
計	299	31.3	18.1

※記号が異なる群は5%水準で有意な差がある

セグメントごとの身体的フレイルの割合比較

【検証①】

- 提案手法による擬人化データを作成したのちに、各人の身体的フレイルの有無を判定し、生データから得られたクラスタと同様の違いが確認されるかを検証

【検証②】

- **<u>生データの時点で各人の身体的フレイルの有無が明らかな場合</u>**を仮定し、生データから得られたクラスタと同様の違いが確認されるかを検証

セグメントごとの身体的フレイルの割合比較（検証①）

生データの場合に得られたクラスタ間での違いと
同様の傾向は提案手法を用いた場合のクラスタ間では確認されず

- 身体的フレイルの判定：Friedら(2001)の基準[10]の修正版を使用

生データの場合

クラスタ	N数（フレイル判定可能）	身体的フレイル割合
		全年代
S1	155	11.0%[a]
S2	701	13.7%[a]
S3	851	20.9%[b]
S4	246	37.0%[c]
計	1,953	19.6%

提案手法を用いた場合

クラスタ	N数（フレイル判定可能）	身体的フレイル割合
		全年代
C1	51	0.0%[a]
C2	80	3.8%[a]
C3	141	11.3%[a]
C4	27	40.7%[b]
計	299	10.0%

※記号が異なる群は5%水準で有意な差がある

セグメントごとの身体的フレイルの割合比較（検証②）

生データの時点で身体的フレイルの判定が明らかな場合を仮定すると、
生データの場合に得られたクラスタ間での違いと
同様の傾向を提案手法を用いた場合のクラスタ間においても確認

- 身体的フレイルの判定：Friedら(2001)の基準[10]の修正版を使用

生データの場合

クラスタ	N数（フレイル判定可能）	身体的フレイル割合
		全年代
S1	155	11.0%[a]
S2	701	13.7%[a]
S3	851	20.9%[b]
S4	246	37.0%[c]
計	1,953	19.6%

提案手法を用いた場合

クラスタ	N数（フレイル判定可能）	身体的フレイル割合
		全年代
C1	51	10.3%[a]
C2	80	16.2%[b]
C3	141	21.3%[c]
C4	27	43.3%[d]
計	299	19.6%

※記号が異なる群は5%水準で有意な差がある

まとめ

提案手法

- PLSAを繰り返し実行することによる個人のプライバシーを保護した情報の作成

提案手法の評価結果

【PLSAによるクラスタリング】

- 生データ、提案手法による擬人化データ、k-means法による擬人化データの全てのデータセットにおいて**4クラス**が採用（AIC基準より）
- 提案手法を用いた場合に得られた幸福度が**最も高い**・**最も低い**クラスタにおいては、生データから得られた幸福度が**最も高い**・**最も低い**クラスタと同様の特徴を確認

【セグメントごとの健康アウトカム比較】

- **総医療費の平均値**については、生データから得られたクラスタ間と同様の違いを、提案手法を用いて得られたクラスタ間においても確認
- **生データの時点で身体的フレイルの判定が明らかな場合を仮定する**と、身体的フレイルの割合においても生データから得られたクラスタ間と同様の違いを確認

参考文献

[1] 内閣府. (2021). 「科学技術・イノベーション基本計画」（閲覧日2023年1月25日）

[2] Boer, de. B. (2020). Experiencing objectified health: turning the body into an object of attention. Medicine, Health Care and Philosophy, 23, 401-411.

[3] 本村陽一・櫻井瑛一・山下和也・井上恵. (2022). 「生活者ビッグデータからのヘルスケアデジタルツイン構築とヘルスケアサービス支援システム」『人工知能学会全国大会論文集』, 36, 1-4.

[4] 本村陽一. (2014). 「サービス工学におけるビッグデータの活用技術」『日本ロボット学会誌』, 32(10), 878-880.

[5]伊藤伸介. (2009). 「匿名化技法としてのミクロアグリゲーションについて」『熊本学園大学経済論集』, 197-232.

[6] 伊藤伸介・村田磨理子・高野正博. (2014). 「ミクロデータにおける匿名化技法の適用可能性の検証―全国消費実態調査と家計調査を用いて―」『統計研究彙報』, 71, 83-124.

[7] 井出絢絵・川本達郎・山下和也・本村陽一. (2017). 「ミクロアグリゲーションを用いた匿名化による確率的潜在空間意味解析」『人工知能学会全国大会論文集』, 31, 1-4.

[8] 厚生労働省. (2022). 「孤独・孤立対策の重点計画」（閲覧日2023年1月25日）

[9] Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 51(2), 211-218.

[10] Fried, P. L. et al. (2001). Frailty in Older Adults: Evidence for a Phenotype. The Journals of Gerontology, 56(3), 146-157.