コンジョイント調査における機械学習を用いた熟慮回答と不適切回答の分類

社会データ構造化センター特任准教授　稲垣 佑典

健康リスク意識に関するコンジョイント調査

◆『コンジョイント分析を用いた健康リスク意識に関するWEBアンケート調査（第2次調査）を実施（「**コンジョイント調査**」と呼称）』

➢ COVID-19に関連した健康リスクについて重視される要因と人々の意識・態度、行動の関連を把握することが目的

➡Web調査の特性を生かし、コンジョイント分析項目を設置

◆ コンジョイント分析とは

➢ いくつかの水準を有する複数の要因で構成された対象（商品・事象）を評価させ、構成要因の好ましさを推定する分析方法（および調査法）

➡要因と水準の組み合わせを変えたコンジョイントカードを繰り返し提示

➡要因・（3または2）水準を組み合わせた状況に対して行動を自粛するか「0：絶対に自粛しない」～「10:確実に自粛する」の11段階で評価

➡提示カードは**全部で18枚**あり、順序効果を回避するため**ランダマイズ**

➢ 第2次調査では「**自己シナリオ**（この状況であなたはどうすべきか）」と「**他者シナリオ**（このような状況の人はどうすべきか）」という2パターンの比較実施

【問題】 類似の事柄を18回反復して尋ねるため負担が大きく、サティスファイシングによる不適切回答が懸念される

◆ サティスファイシング/サティスファイスとは

➢ 対象者が必要な認知資源を割かず、最善でなく満足できる程度の回答行動をとること（端的にいうと“手抜き回答”） e.g., Krosnick (1991); Tourangeau, Conrad & Couper (2013); Maniaci & Rogge (2014); 三浦・小林(2015)

➡コンジョイント項目では他項目より**ストレートライナー回答割合**が増加

➡さらに不適切回答か熟慮の末の回答か不明であるという問題発生

➢ 対策としてIMC(Instructional Manipulation Check)、DQS(Directed Questions Scale)をはじめとした検知項目（screener）を導入（付録1参照）

【研究目的】 検知項目と機械学習の手法を使い、不適切回答と熟慮回答の判別と別調査（第1次調査データ）における予測を試みる

コンジョイント調査概要

◆ 調査設計

1）調査対象：全国の20～64歳の男女（都道府県の性・年代別人口構成比で割当て）

2）計画サンプルサイズ：2,000名×2（自己/他者シナリオの2グループ）

謝辞 本調査は情報・システム研究機構未来投資型プロジェクト『要因実験的な社会調査を用いた健康リスクに関する情報提供と行動変容の関連性の研究：新たな科学コミュニケーションモデルの創出（代表: 加藤直子）』の一環として実施した

◆ 回収結果

1）調査時期：2021年12月8日~11日

2）有効回収サンプル：**自己シナリオ** 2,707名，**他者シナリオ** 2,754名

（※参加依頼数不明で回収割合算出不能）

3）調査会社側の無効条件：①登録情報と性・年齢・居住地不一致、②極度の短時間回答（基準不明）➡71サンプルを事前に除外して納品

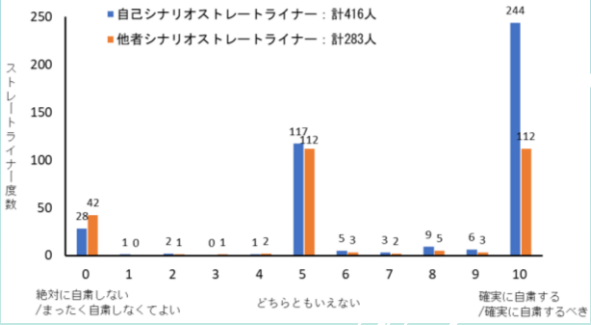

図1 コンジョイント項目における自他シナリオ別ストレートライナー度数

クラスタ分析とSVMよる回答傾向とストレートライナーの分類

◆尋ね方のフレームが異なると、ストレートライナー（SL）にも違い発生（図1 参照）

➡不適切回答SLの発生は尋ね方による影響を受けないはずなので、違いがあるということは、SLに熟慮回答が紛れ込んでいる可能性を示唆

不適切回答者除外と仮説（付録2）検証のため、SLを分類する必要がある

➢ 第2次調査の検知項目に**非階層的クラスタ分析（K-means++法）**を施し、**不適切回答SL/熟慮回答SL**の区分を試みた

➡エルボー法（付録3 図2）で自己シナリオ/他者シナリオともに2つのクラスタ導出し、反応の比較（付録4 表1）から不適切回答と適切回答クラスタと解釈

➡SLと検知項目のクラスタ分析の不適切回答/適切回答クラスタをクロスさせて、不適切回答SL/熟慮回答SLに区分（付録5 図3）

✓ 自己/他者シナリオともに**「0」と「10」は熟慮SLの割合多い**が、**「5」は不適切SLの割合多く**中間回答は不適切回答が多いことが示唆

◆検知項目を導入しなかった第1次調査でも不適切回答SLと熟慮回答SL識別を行いたい！

➢ 第2次調査の**非SL/不適切回答SL/熟慮回答SL**という目的変数に対し、第1次・2次共通かつ2次調査のクラスタ分析で違いが顕著な特徴量(付録6 表2)を使った**分類目的の機械学習：SVM（Support Vector Machine）**を用いて検知項目の無い調査データでのSLの予測を試みた

➡**非SL：0、不適切回答SL：1、熟慮回答SL：2 をカテゴリカルな目的変数**とした多値分類非線形SVM

- 欠損値は平均値で置き換え、データは標準化し、目的変数の不均衡データ状態解消のためSMOTE（Synthetic Minority Over-sampling Technique）によるオーバーサンプリングで少数カテゴリの拡張をする加工を実施

【分類結果】

➢ **自己シナリオのAccuracyは68%、他者シナリオは約63%**となった（精度指標とハイパーパラメータの設定は付録7 表3参照）

➡尾崎・鈴木（2019）の機械学習モデルでは最高で55.6%の正答率、Pei et al.(2020)のディープラーニングでは　75.65%の正答率であり、一定以上の精度が確認できた　※検知項目の使用をしていない尾崎・鈴木と比べて**検知項目を使用した学習の方が、予測精度高い**（調査会社は嫌がるが有効）

第1次調査データを使用した不適切回答者の予測

◆自己シナリオのSVM学習器は特徴量が第1次調査データ（付録8 調査概要）と共通しているため、これを用いて**第1次調査データのコンジョイント項目におけるストレートライナーの予測分類が可能**➡結果は図2（拡大図は付録8 図4）の通りとなった（非SLと予測された明らかな**誤分類は117名で、29.3%**ほど）

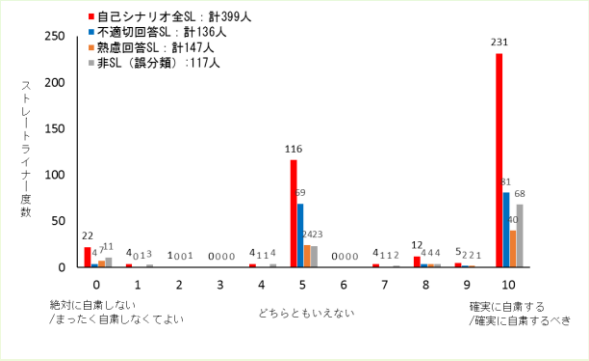

図2 SVMおける自己シナリオにおけるストレートライナー分類予測結果

【結果のまとめ】

◆自己シナリオと他者シナリオともに、SLが0・5・10でSLが発生する傾向が見られた

➢ クラスタ分析の結果、自他シナリオともに**5においては不適切回答SLの割合高い**

➢ クラスタ分析の結果、自他シナリオともに**10においては熟慮回答SLの割合高い**

➢ クラスタ分析の結果。他者シナリオでは、**0の熟慮回答SLの割合が高い**

➡他者への行動強制へのためらいが生じていたという仮説が支持

➡コンジョイント分析ではシナリオのフレーム次第で回答傾向が変わることが示唆

◆SVMによる学習の精度は自己・他者シナリオともに60%以上で、それなりの結果

➡先行研究よりも良い精度となり、検知項目（screener）と機械学習を併用することの有用性を示唆

➡デフォルトパラメータのため、**パラメータチューニングで精度向上が見込める**

◆SVMの学習器から第1次調査のSLの分類予測では、**熟慮回答SLの割合少なかったが、誤分類である非SLに吸収された可能性がある**

付録1

（クラスタ分析に使用した）第2次調査サティスファイス検知項目概要

- IMC

以降の質問は皆さまに間の内容をしっかりとお読みいただいたうえで、想定されている状況を想像しながらお答えいただく必要があります。
そこで大変恐縮ではございますが、まずは皆さまがここでの説明をお読みいただいたか、確認させていただきたく存じます。

説明をお読みいただいた方は、本ページ下部にある「質問の答え方について　理解した/理解していない」というラジオボタンには何もチェックを入れずに、「次へ進む」ボタンだけを押して次のページにお進みいただきますよう、お願いいたします。

注意書き

質問の答え方について・・・

1　理解した
2　理解していない

ダミーの回答選択肢（チェックを入れると、1度目のみ警告出て再度説明を読むよう促す）

次へ進む

←指示ではこのボタンだけ押すように要求

➡指示遵守(違反0)、1度の警告で通過（1回違反）、指示違反(2回違反)

- DQS×2（ここでは〇〇という選択肢を選んで下さい）

- 矛盾（あり得ない）項目：過去1週間で行なった活動を選んで下さい
  - ✓ インターネット利用（本調査への参加も含む）←選ばないと矛盾
  - ✓ 東京オリンピック生中継←調査は2021年12月実施のため選ぶと矛盾

- コンジョイントカード内容確認項目×2：先ほどのカードの感染者数は？

その他

- 低認知欲求尺度7項目(7点尺度)合計点

- 総回答時間

上記変数はクラスタ分析する際に
欠損値を平均値で置き換え、事前に標準化を施した

付録2

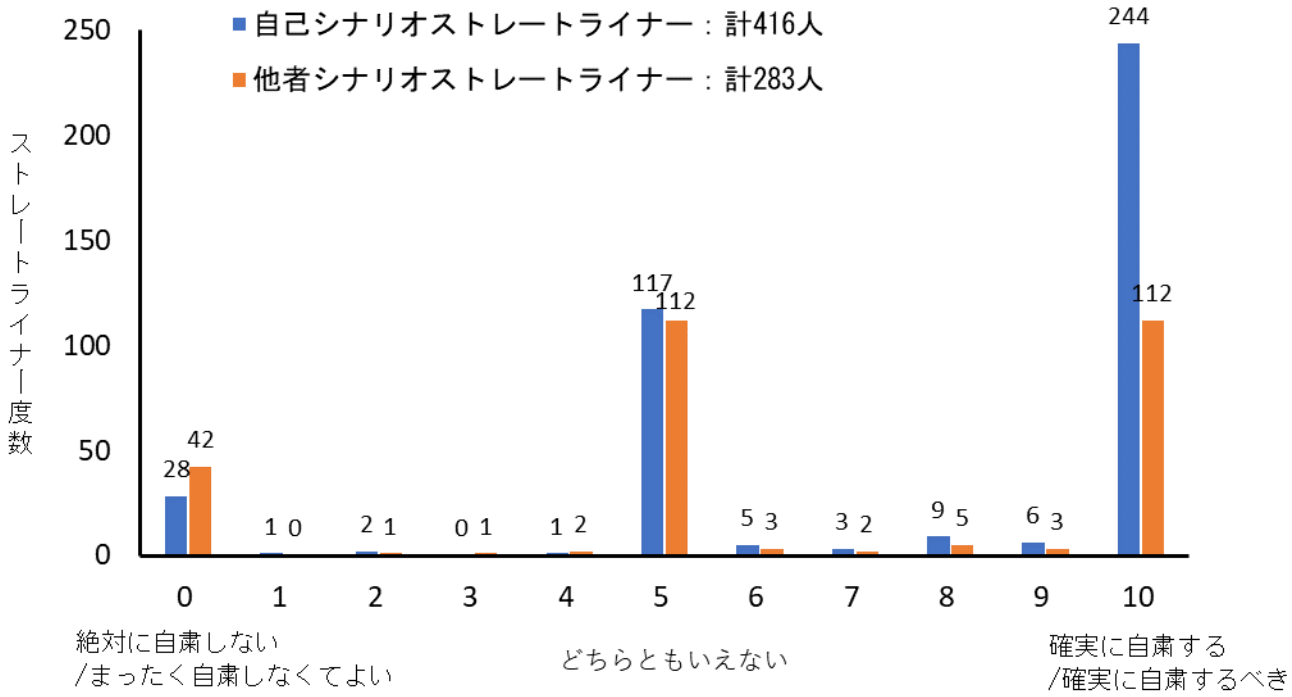

図1 コンジョイント項目における自他シナリオ別ストレートライナー度数（拡大）

- 両シナリオともに0・5・10という“キリ”のよい選択肢でストレートライナーが頻発

- 自己シナリオ：10（確実に自粛する）が他者シナリオより多い
- 他者シナリオ：0（まったく自粛しなくてよい）が自己シナリオより多い

◆ 尋ね方のフレームが異なると、ストレートライナーに違いが生じた

➡不適切回答のストレートライナーは尋ね方を気にしないため違いはないはず
熟慮した末の回答が紛れ込んでいるのでは？

仮説

- 自己シナリオでは、リスク回避傾向が強まり、少しでもリスクあれば「10」を選択するように
- 他者シナリオでは、他者への行動制限強要へのためらいにより「0」の選択が生じやすい

検証のためには不適切回答のストレートライナーの識別が必要

付録3

ヒジとなるクラスタ数は2か8？
（※シンプルな2を採用）

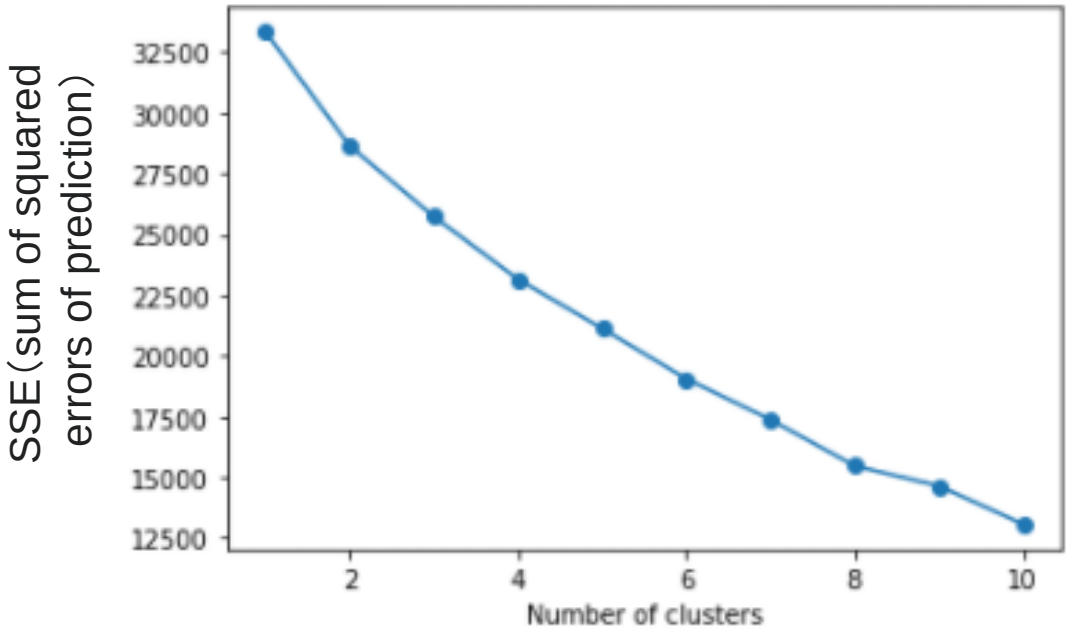

図2.1 エルボー法によるクラスタ数の決定（自己シナリオ）

やはり2か8？
（※自己シナリオと同じ2を採用）

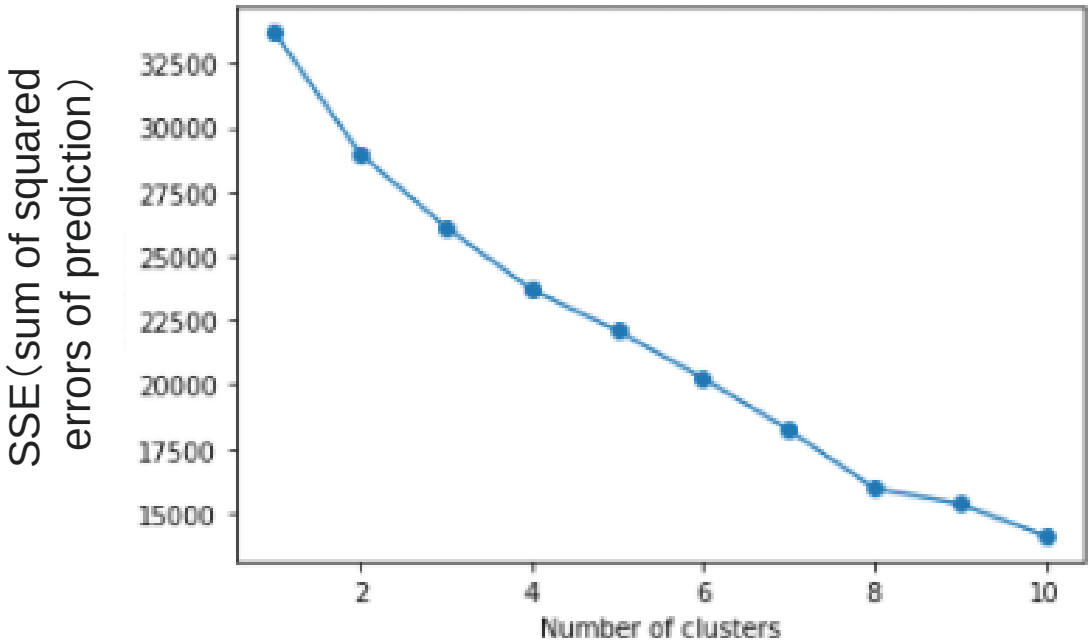

図2.2 エルボー法によるクラスタ数の決定（他者シナリオ）

付録4

クラスタ分析で導出されたクラスタ解釈のための分析

表1 クラスタ別にみた分析変数への反応の比較結果

クラスタ分析の分類：	クラスタ1（不適切回答）	クラスタ2（熟慮回答）	χ^2/t	p
	度数/平均値			
自己シナリオ：				
IMC非通過ダミー:1	477	689	514.78(df=1)	0.001
（２回違反） :0	92	1520		
DQS1非通過ダミー:1	189	5	758.06(df=1)	0.001
:0	380	2204		
DQS2非通過ダミー:1	430	717	346.93(df=1)	0.001
:0	139	1492		
矛盾回答1ダミー:1	24	33	16.71(df=1)	0.001
（選ぶと矛盾） :0	545	2176		
矛盾回答2ダミー :1	441	89	1582.17(df=1)	0.000
（選ばないと矛盾） :0	128	2120		
CJ確認1非通過ダミー:1	414	1054	113.89(df=1)	0.001
:0	155	1155		
CJ確認2非通過ダミー:1	418	1001	143.46(df=1)	0.001
:0	151	1208		
低認知欲求尺度得点(平均値)	4.63	3.78	10.22(df=998.65)	0.001
総回答時間(秒)	735.32	1450.47	4.56(df=1913.82)	0.001
他者シナリオ：				
IMC非通過ダミー:1	526	689	578.08(df=1)	0.001
（２回違反） :0	87	1506		
DQS1非通過ダミー:1	167	8	592.40(df=1)	0.001
:0	446	2187		
DQS2非通過ダミー:1	457	741	325.97(df=1)	0.001
:0	156	1454		
矛盾回答1ダミー:1	39	37	37.80(df=1)	0.001
（選ぶと矛盾） :0	574	2158		
矛盾回答2ダミー :1	451	89	1490.90(df=1)	0.000
（選ばないと矛盾） :0	162	2106		
CJ確認1非通過ダミー:1	454	979	166.43(df=1)	0.001
:0	159	1216		
CJ確認2非通過ダミー:1	448	908	193.04(df=1)	0.001
:0	165	1287		
低認知欲求尺度得点(平均値)	4.22	3.75	9.29(df=1029.48)	0.001
総回答時間(秒)	646.00	1653.71	6.47(df=2567.30)	0.001

付録5

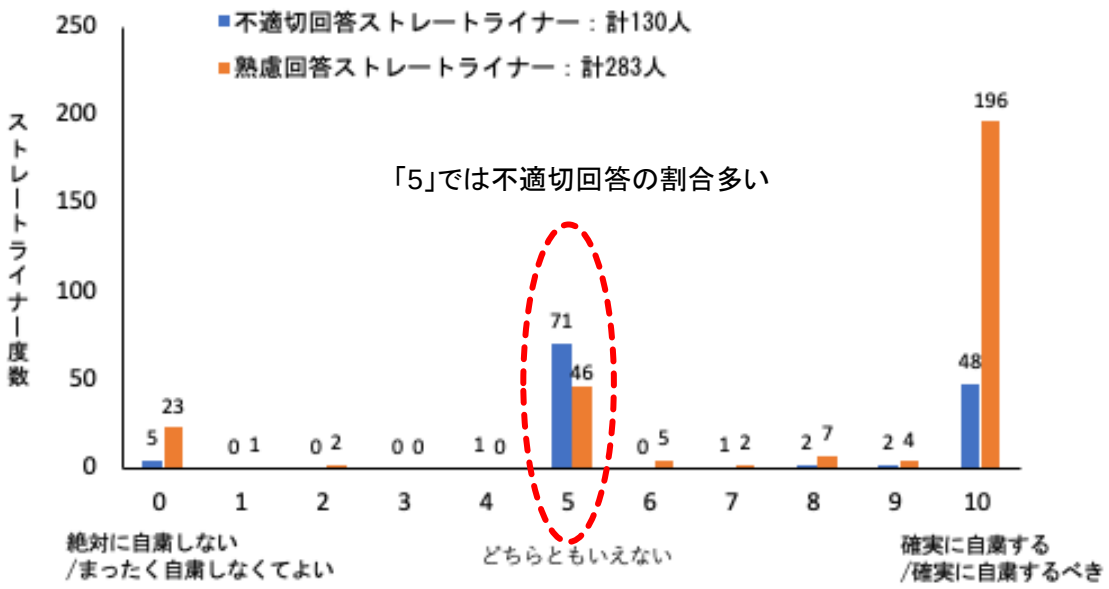

図3.1 クラスタ分析結果にもとづくストレートライナー再分類後の分布（自己シナリオ）

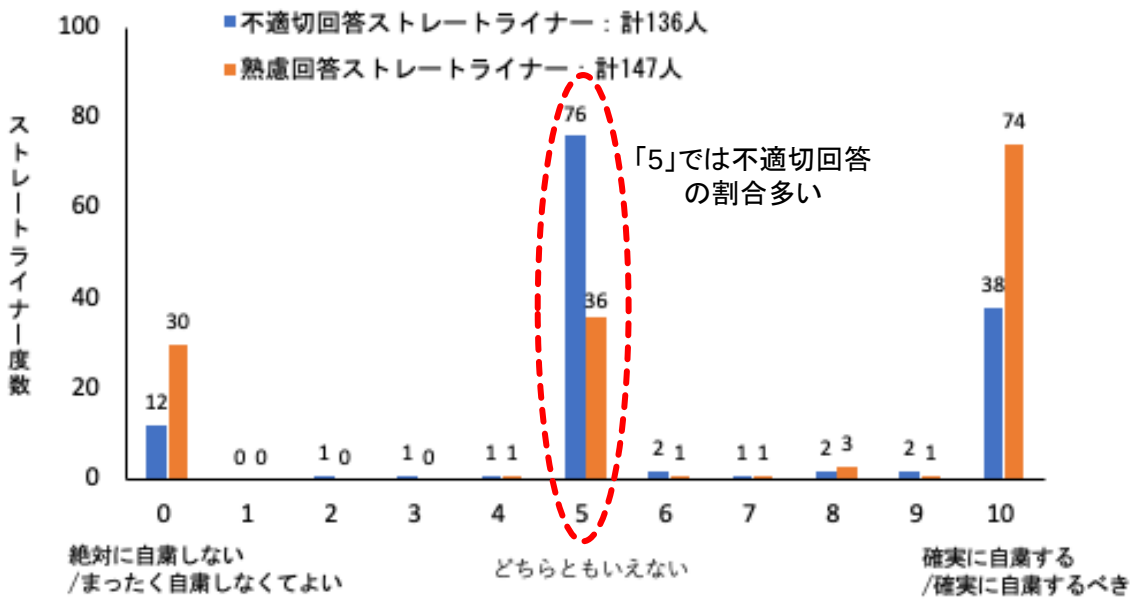

図3.2クラスタ分析結果にもとづくストレートライナー再分類後の分布（他者シナリオ）

付録6

SVMの学習・予測の特徴量に使用した第1次・2次調査共通項目概要

- TIPI-J（日本版Ten Item Personality Inventory）ストレートライナーダミー
  ➡小塩ら（2012）の簡易版Big5尺度（10項目、7点尺度）
  ➡5性格特性ごとに2つの正順・逆転項目で構成されており、ストレートライナーでは、回答内容の矛盾大きい

- Covid-19高リスク同居者有無10項目Yesストレートライナーダミー
  ➡10項目にわたる高リスク者（高齢者、基礎疾患保有者、免疫抑制剤接種者 etc…）との同居有無を尋ねた質問
  ➡すべてにYesが該当する対象者は極めてまれ

- 感染予防行動項目（5項目、5点尺度）平均点
- 感染回避行動項目（3項目、5点尺度）平均点
  ➡不適切回答者と思われる人々とそうでない人々との間で回答に差が確認

- ニューメラシー尺度（10項目）正解得点
- 科学リテラシー尺度（11項目）正解得点
  ➡両尺度とも稲垣・加藤・前田（印刷中）および前回発表で不適切回答と強い関連性が示唆

- 調査協力理由（10項目、無制限連記式）合計数
  ➡稲垣・加藤・前田（印刷中）で調査協力理由と不適切回答の間に関連性が示唆

クラスタ分析の分類結果で比較しても多くの変数に差があり、特徴量として有効

表2 SVMの特徴量とする変数のクラスタ分析を基にしたSL比較結果

クラスタ分析に基づくSLの分類：	不適切回答SL	熟慮回答SL	χ^2/t	p
	度数/平均値			
自己シナリオ：				
TIPI-J_SLダミー:1	67	34	76.43(df=1)	0.000
:0	63	252		
リスク者同居_YesSLダミー:1	3	0	6.65(df=1)	0.010
:0	127	286		
感染予防行動平均点	2.19	1.47	6.16(df=157.57)	0.000
感染回避行動平均点	3.95	4.22	2.68(df=206.08)	0.010
ニューメラシー得点	1.53	5.55	14.23(df=314.33)	0.000
科学リテラシー得点	1.85	5.56	12.82(df=294.20)	0.000
調査協力理由数	1.7	2.67	6.61(df=222.84)	0.000
他者シナリオ：				
TIPI-J_SLダミー:1	59	29	18.45(df=1)	0.000
:0	77	118		
リスク者同居_YesSLダミー:1	0	0	—	—
:0	136	147		
感染予防行動平均点	1.76	1.69	0.73(df=265.10)	4.660
感染回避行動平均点	3.70	4.07	3.11(df=256.52)	0.020
ニューメラシー得点	2.80	5.62	7.22(df=278.84)	0.000
科学リテラシー得点	2.71	5.43	7.09(df=281.00)	0.000
調査協力理由数	1.64	2.57	5.72(df=246.70)	0.000

付録7

表3 非線形SVMの精度指標とハイパーパラメータの設定

（3クラスマクロ平均※）	Accuracy（正解率）	Precision（適合率）	Recall（再現率）	F1-score
自己シナリオ：SL3分類	68.00%	67.96%	68.09%	67.92%
他者シナリオ：SL3分類	62.83%	63.35%	63.24%	62.39%

Accuracy：正しく予測されたサンプルの割合（の3クラス平均）
Precision：陽性と予測した中で、実際に陽性だった割合（の3クラス平均）
Recall：実際に陽性のものの中で、正しく陽性と予測できた割合（の3クラス平均）
F1： PrecisionとRecallの調和平均（の3クラス平均）

※1 ストレートライナーの不均衡データを事前にSMOTE（Synthetic Minority Over-sampling Technique）でリサンプリングして補っており、サンプルサイズに偏りがないことから重みを考慮しない**シンプルなマクロ平均**（各クラスの指標を単純平均した値）を算出
※2 全データ中、学習に75%を使用し、テストには残り25%を使用

	C	Krnel	gamma
ハイパーパラメータ	1.0	RBF	0.01

C：誤分類を許容する程度のパラメータ、大きいほど誤分類を許容しない
RBF(Radial Basis Function kernel)：非線形の分類アルゴリズム
Gamma：RBFの複雑さを規定するパラメータ、大きいほど複雑な分類になる

今回はデフォルト設定のため、今後パラメータチューニングで精度向上が見込める

付録8

第1次調査概要

◆ 調査設計

1）調査対象：全国の20～64歳の男女（都道府県の性年代別人口構成比で割当て）

◆ 回収結果

1）調査時期：2021年3月22～25日

2）有効回収サンプル：2,731名（参加依頼数は不明で回収割合は算出不能）

3）調査会社側の無効条件：

①登録情報と性・年齢・居住地不一致

②極度の短時間回答（基準不明）

③マトリクス形式項目への回答がすべてストレートライナー

➡83サンプルを事前に除外して納品

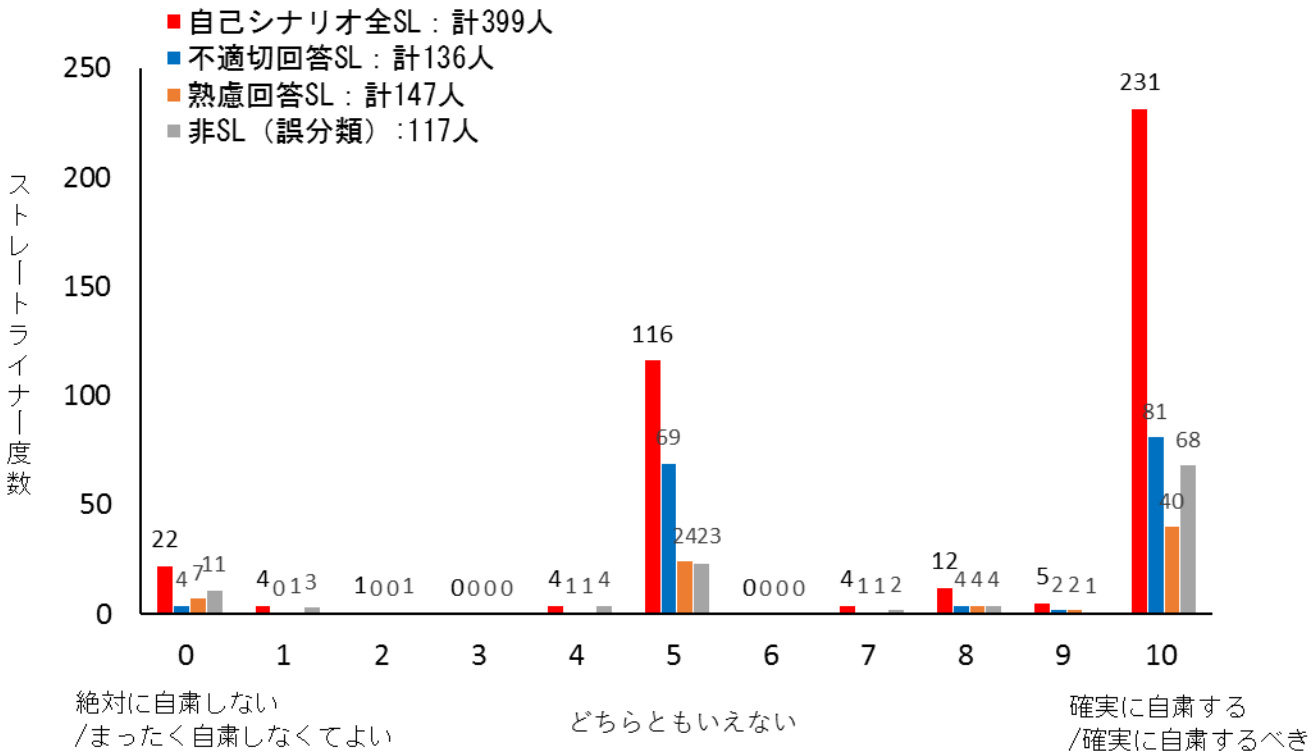

図4 SVMおける自己シナリオにおけるストレートライナー分類予測結果

◆ クラスタ分析の結果と異なり熟慮回答SLに分類された数が少ない

➡誤分類である非SLに吸収されている可能性がある