クラスタリングによる自由記述回答の要約と選択肢回答空間に射影による解答群間の連関の可視化

根本颯汰 [1,2] 藤本一男 [1,3]
[1] 国立研究開発法人 情報通信研究機構（NICT）サイバーセキュリティ研究所
ナショナルサイバートレーニングセンター
[2] 法政大学大学院 理工学研究科応用情報工学専攻 [3] 津田塾大学 数学・計算機科学研究所
{kazuo.fujimoto, sota_nemoto}@nict.go.jp kazuo.fujimoto2007@gmail.com

概要

本研究では，非階層型クラスタリングによって自由記述回答を「要約」，それを構造化データ解析（MCA/SDA）の手法で生成された選択肢回答空間に射影し，この二つの回答群の連関を幾何学的に可視化する手法を開発した. 本研究では，自由記述回答を非階層型クラスタリングによって分類し，そのクラスタ番号を，選択肢回答のデータフレームに接続し，多重対応分析における追加変数として処理した. この処理によって，選択肢回答によって生成された変数/個体空間上にクラスタ番号として「要約」された自由記述回答をプロットすることが可能になる. こうして，自由記述部分の分析条件が拡大することを示した.

1 はじめに

多くの調査票調査には, 設問は選択肢が用意された設問（構造化された設問）と自由記述のような非構造化設問が用意されている. 構造化された設問は, 回答する側を質問する側の枠にはめることであり, 分析基準は明確ではある. しかし, その反面, 実際に回答者に生じている変化を捕捉できないという「硬直」した側面も有している. そうした場合, 回答選択肢に「その他」が設けられ,「その他」の内容を自由記述で回答するように促されるが, そこでは「自由」ゆえの分析上の困難が生じる. そのため, ある時期（テキストマイニングの普及）までは「分析できない自由記述はとるな」という方針も存在していた. 今日では,KH-Coder[1)] [1] のようなテキスト・マイニングツールの普及, また R や Python でのテキスト・マイングのパッケージ, モジュールの普及により利用するハードルは低くはなっている. しかし, これらのツールを用いることで自由記述部分自体の分析は容易になったとはいえ, その分析内容と他の選択肢回答との連携は非常に限定されたものになっている[2)]. こうした状況に鑑み, 本研究では, テキストマイニング, クラスタリングと多重対応分析を組み合わせ, 以下のような分析方法を実現した (図 1).

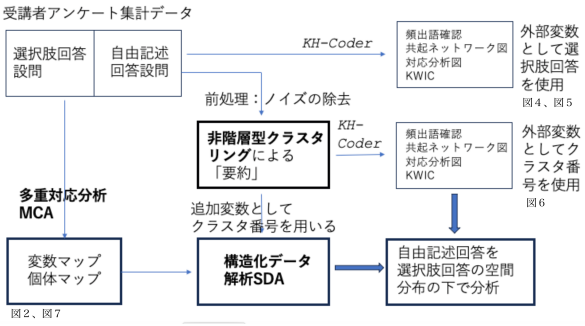

図 1 MCA によって生成された変数マップ

- 自由記述回答を, 機械学習を用いてクラスタリングする.
- クラスタリングによって「要約」された自由記述回答を回答者ごとに選択肢回答のデータフレームに追記する. これを, 多重対応分析をもちいた構造化データ解析の手法をもちいて, 回答空間にプロットする.

こうすることによって, 選択肢回答と自由記述回答の連関が幾何学的に可視化できた.

分析対象としたデータは, 筆者たちが在籍する NICT が主催する CYDER[3)] の受講者アンケートの集

1） https://khcoder.net/

2） KH-coder の共起ネットワーク図, 対応分析図で投入可能な外部変数は 1 変数である.

3） CYDER（Cyber Defense Exercise with Recurrence：実践的サイバー防御演習）とは，サイバー攻撃を受けた際の一連の対応（インシデント対応）をパソコンを操作しながらロールプレイ形式で体験できる演習 [2]

表 1　調査票のうち空間生成に選択した変数の度数分布表

スキル向上	度数	理解	度数	講師説明	度数	サポート適切	度数	対応適切	度数
非常に向上したと思う	367	よく理解できた	559	とてもわかりやすかった	755	とてもそう思う	647	とてもそう思う	695
向上したと思う	1553	理解できた	1032	わかりやすかった	1019	そう思う	1276	そう思う	1188
変わらない	33	理解できない内容があった	352	ふつう	192	あまりそう思わない	57	あまりそう思わない	32
わからない	36	理解できない内容が多かった	53	わかりづらかった	18	まったくそう思わない	3	まったくそう思わない	3
NA	12	*NA*	5	とてもわかりづらかった	5	*NA*	18	質問はしなかった	72
合計	2001			*NA*	12			*NA*	11

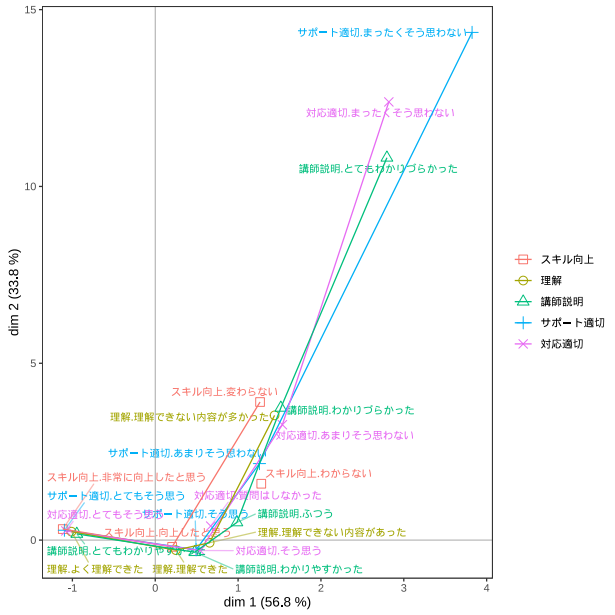

図 2　MCA によって生成された変数マップ

計データを使用した.

2　多重対応分析 MCA によるデータの構造化

2.1　設問構造

多重対応分析 (MCA) で回答空間を生成するために投入する変数は, 受講者アンケートのうち,「理解できたか」「説明はわかりやすかったか」「受講してスキル向上したか」「サポートは適切であったか」「(演習の) 進度は適切であったか」の 5 変数である. この 5 変数をもちいて, 回答空間 (変数空間と個体空間) を生成した[4].

2.2　追加変数による回答空間分析

追加変数 (Supplementary Variables) は, 空間の座標軸生成には影響を与えないが, 座標値をもつことが可能とする変数処理の方法であり [3], これを回答空間の構造分析に用いることができる. 例えば, 年代 (年齢) を追加変数として射影すれば, 回答空間の年代構造が, 業務経験年数を射影すれば, 業務経験年数構造が可視化される.

その追加変数として, 自由記述設問で生成されたクラスタ番号をプロットすることで, 回答空間における「分類・要約」された自由記述の空間配置が可能になり, それから, 回答空間の構造を分析することが可能になる.

2.3　生成された空間の解釈

変数空間と個体空間は, 同じ慣性 (分散) の座標軸で表現されるので, 座標軸の解釈は, 変数空間において行うことになる. まず, 原点は, 全体の平均の位置を示していることを確認する. そのため, 軸であれ点であれ解釈する場合には原点との関係, そして座標軸に対してどちら側であるかに注目して解釈していく. 生成された空間の座標軸が体現している慣性 (分散率) は次のようになっている.1 軸で 56.8 %,2 軸までで 90.6 %,3 軸まで含めると 97.8 %である. ここでは 1 − 2 軸までを分析対象とし,10 %は誤差として扱うことにする.

2.4　自由記述回答の「要約」

クラスタリングによって分割対象とする自由記述回答は, 以下のものがある.1) 設問「理解」を選択した理由,2) 受講して得られた気づき,3) 全体的感想, である. 内容を検討した結果,CYDER の中心プログラムである集合演習に直接関係するものとして,1) を分析対象とする. これを選択肢解答による個体マップ [注] にプロットすることで, グルーピングされた自由記述解答を選択肢回答空間における配置として解釈することが可能になる.

4)　生成された変数空間については図 2, 個体空間は, 図 3 を参照のこと. また MCA の基本的考え方, とくに本稿で用いる, 追加変数を用いた構造化データ解析の考え方については文献 [3] を参照のこと.

表 2　コースと自由記述回答ごとのクラスター数と要約

自由記述	クラスター数	要約例
理解度理由	6	講義形式 / 複数回受講 / インシデント対応 / テキスト・解説, etc...
気付き	6	講義時間 / 講義内容 / チューター / 感想 / 特になし / 特にありません
全体感想	2	改善点 / 特になし・感謝コメント

表 3　理解度理由におけるクラスターの要約と回答例

要約	回答例
講義の実施形式	・ハンズオンを交えたグループ形式でわかりやすく記憶にも残りやすい講義だった。 ・流れに沿って一つづつ課題をこなしていくことで身についたように感じた。
繰り返しの受講	・2 回目の研修ということもあり理解を深めることができたと思う。 ・繰り返し受講しているため徐々に知識が定着してきたと感じる。
テキスト・解説	・テキストも講師の方の説明も丁寧であった。 ・IT 知識のない人でも理解できるように丁寧に説明していただけたのでよく理解できた。

3　非階層型クラスタリングによる自由記述回答の「要約」

自由記述回答を回答空間に射影し, 選択肢回答との連関の可視化を行うために, 我々は非階層型クラスタリングによる自由記述回答のクラスターの作成とクラスタリングの結果と TF-IDF によるクラスターごとの要約を作成した. クラスタリングの対象と今後の拡張性の 2 つの観点から, 非階層型クラスタリングを採用した. 具体的な理由として, 今回のクラスタリングの対象は自由記述回答であり, 回答はフォーマットが定められていないため, 階層構造による分類は困難であり, また今回は 2020 年度に実施された A コースのデータを対象に分析を行ったが, 将来的には他年度や他コースのデータも交えた分析を目指しているため, データが大量に増えてもクラスタリング精度の落ちにくい傾向がある.

定量的に意味を持ち分析に有意義なクラスターや適切なクラスター数を定義するために, 我々はクラスタリングに X-means[4] を用いた. これは互いに近いデータ同士は同じクラスタに属するという考えを基に, データを k 個にクラスタリングする教師なし学習の一種である K-means[5] をクラスター数 k を自動的に定めるように拡張された手法である.

クラスタリングを行う前に, 分析対象である,1) 設問「理解」を選択した理由,2) 受講して得られた気づき,3) 全体的感想, の 3 つの自由記述回答に対して前処理をした. 前処理は, 受講者による表記揺れや時間内に回答を書き切れなかった回答をなるべく減らすために,1) 句読点, 空白の削除,2)「よかった」や「特になし」などの特定の表現を除いた 5 文字以下の回答の削除, の 2 つを実施した. 前処理を行った回答に対して, 日本語の wikipedia のコーパスのによって事前学習された RoBERTa[6] を用いて, 回答の 768 次元の埋め込み表現を取得した. RoBERTa[6] は文書分類タスクなどで効果が示されている機械学習モデルであり, 我々の目的は埋め込み表現を取得するのみであるため, 公開されているモデル[5] を読み込み, 追加学習はしていない. 取得した埋め込み表現から K-means++[7] によって各表現の重み付き確率分布を考慮して初期値とを適切に設定した後に,X-means[4] によって 1 から 50 個の範囲でクラスター数を探索し, クラスタリングを行った.

最後にクラスタリングされた自由記述回答に対して, 回答内に含まれる名詞や形容詞による TF-IDF の結果とクラスタリングされた回答を目視によって確認し, 人の手によって要約を作成した. X-means による自由記述回答に対するクラスター数とクラスターの要約を表 2 に示す. クラスタリングの導入によって, 自由記述回答の一定の分類が可能になり, 実際に, 受講して得られた気づきに対しては「もっと講義時間を増やしてほしい」,「進行スピードが遅たった」などの “講義時間” に関する気付きや「チューターの方のアドバイスがわかりやすかった」などの “チューター”,「実践的でわかりやすかった」などの “感想” や「特になし」「特にありません」などのクラスターごとの特徴が確認できた. “感想” や “特になし” などの講義の改善のために有意義でない回答に対するクラスタリングは明確に分類することが可能だったが, より詳細な “講義形式” や “講義内容” に関する回答に関しては分類性能は劣り, 課題点であ

5)　https://huggingface.co/nlp-waseda/roberta-large-japanese-seq512

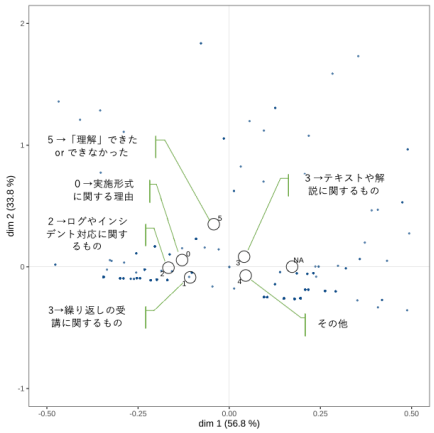

図 3　MCA によって生成された変数マップ上へのクラスタの plot

る. また, 今回の要約の作成方法はヒューマンエラーを減らすために自動化できるのが望ましいため, こちらも我々の今後の課題点である.

4　構造化データ解析による「要約」の可視化

3 節で記述したように, 自由記述回答をクラスタリングし, それに名前をつけた. それを選択肢回答の空間先生をおこなったデータフレームに接合して, それを追加変数（サプリメンタリ変数）として plot する[6]に記されている. こうすることで, その要約された自由記述を回答空間に位置付けることが可能になる.

ここに掲載したのは, 設問理解に付随してつくられた「理解の回答」を選んだ理由を聞いているものである. 取得されたクラスタは以下のように命名された.

- 0 →実施形式に関する理由
- 1 →繰り返しの受講に関するもの
- 2 →ログやインシデント対応に関するもの
- 3 →テキストや解説に関するもの
- 4 →その他
- 5 →「理解」できた or できなかった
- （無回答は NA になっている）

4.1　外部変数を投入した共起ネットワーク

ここで,KH-coder によって, この自由記述回答部分,「理解」を回答した理由を処理し外部変数にこのクラスタ番号を投入したものをみてみたい. 1）KH-coder で, 理解理由の共起ネットワークをかかせると, 語と語の共起関係が可視化される (図 4). この図は「自由記述回答」の内部の構造を探査している. 2）次に外部変数を投入する. まず, 外部変数として設問「理解」を投入する. これによって, 設問「理解」の回答との関係は可視化できる（図 5）. ただ, それはあくまでも, 設問「理解」との限定された関係である.（1 もそうだが, これはこれで, 自由記述回答の重要な情報ではある.）3）そして, この外部変数を 3 節で解説したクラスタリングの結果であるクラスタ番号を投入して実施した（図 6）.

4.2　クラスタの空間的配置

こうして手に入れた共起ネットワーク図に加えて, 図 3 で示した, 回答選択肢によって生成された空間内でのクラスタ番号のプロットがある. この回答空間の座標は, 原点が全体の平均位置, 水平軸の左側が「理解」に限らず積極的・肯定的な回答に対応してた. また, 水平軸の右側の方向が, 消極的・否定的な回答に対応している. これらを踏まえて多重対応分析のマップにプロットしたクラスタ番号と KH-coder の外部変数としてクラスタ番号を投入した共起ネットワーク（KH-coder の対応分析図も含めて）を（KWIC 表示機能も援用しながら）見ていくことになる.

選択肢回答の座標軸の左方向（肯定的・積極的）に, クラスタ番号,2,0,1 が並び（これらは演習の内容に言及している）, より全体の平均に近く, 垂直軸上側（否定的・消極的）方向に 5（「理解」をめぐるもの）があることが確認できる. さらに, 原点の右側, 垂直軸のすこし上向の位置に 3（テキストや解説）が位置していることがわかる. このように, 自由記述部分を分解, 要約し回答空間にプロットすることで肯定的 – 否定的, 積極的 – 消極的な傾向を具体的に可視化することが可能なった.

5　おわりに

以上みたように, 非構造化データである自由記述回答を, クラスタリングすることによって「要約」が取得でき, その幾何学的な位置を選択肢回答空間内に確認することができた. これをもって, 日々変化していく情報セキュリティ領域のトレーニング手法の改善に寄与することを目指していく.

6）こうした分析手法を構造化データ解析（SDA）と呼ぶが, 分析の中心に多重対応分析 MCA をもちいるデータ分析手法である, 幾何学的データ解析の一パートである. 詳細は, 文献 [3]

謝辞

本研究にあたり,CYDER 実施の中心センターである NICT ナショナルサイバートレーニングセンターの園田道夫センター長, 同サイバートレーニング研究室の花田智洋室長には, データ利用を含め, さまざまな便宜を図っていただきました. また, 中川哲也氏, 阿部則夫氏にはアンケートデータの構成,CYDER のシナリオ, コースプログラムに関する質問に対応していただきました. 記して感謝いたします.

参考文献

[1] 樋口耕一. 社会調査のための計量テキスト分析-内容分析の継承と発展を目指して、第 2 版. ナカニシヤ出版, 京都, 2020. OCLC: 1149044718.

[2] NICT. Cyder（cyber defense exercise with recurrence：実践的サイバー防御演習）, 2022. https://cyder.nict.go.jp/.

[3] Brigitte Le Roux, Henry Rouanet, 訳：大隅昇, 小野裕亮, 鳰真紀子. Multiple Correspondence Analysis(多重対応分析). SAGE publisher（オーム社）, 2010(2021).

[4] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In **Icml**, Vol. 1, pp. 727–734, 2000.

[5] James MacQueen, et al. Some methods for classification and analysis of multivariate observations. In **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, Vol. 1, pp. 281–297. Oakland, CA, USA, 1967.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[7] David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In **Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms**, pp. 1027–1035, 2007.

付録

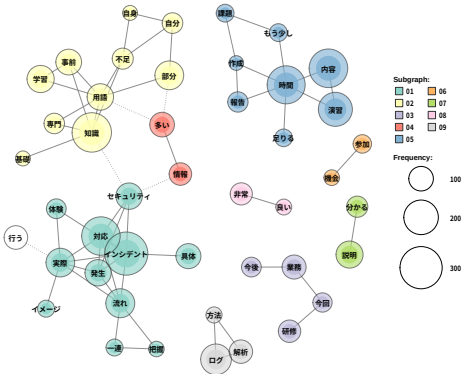

図 4　外部変数なしの共起ネットワーク図

A　多重対応分析による個体マップ

多重対応分析によって，二つのマップが描かれる．変数マップと個体マップである．変数マップと個体マップは，生成される座標軸の体現している分散が同じあるため（これが対応分析の「対応」），軸の解釈は，変数マップでおこなっている．なお，変数マップ（2）は，変数間の関係がみえやすいように，アスペクト比を 1 にしていない．ここに掲載した個体マップは，アスペクト比を 1 にしている．クラスタ番号ごとの平均点のプロットは，中心部分に集中するが，そこには，クラスタの相互関係ははっきりでている．本文には，その部分を拡大したものを掲載しているが，ここには，個体マップ全体をアスペクト比 1 で掲載する（図 3）．

B　KH-Coder による共起ネットワーク

B.1　外部変数なしの共起ネットワーク

この図 (4) で明らかにされるのは，分析対象の自由記述部分内での語の共起関係（相互関係）である．

B.2　変数「理解」を外部変数として投入した共起ネットワーク

これによって，この自由記述回答に関連した変数「理解」の回答カテゴリとの関係が可視化されている（図 5）．

B.3　クラスタ番号を外部変数にした共起ネットワーク

図 6 は，自由記述回答の分析に際して，外部変数として，クラスタ番号を投入した．それによって，各クラスタでの共起関係が可視化されている．

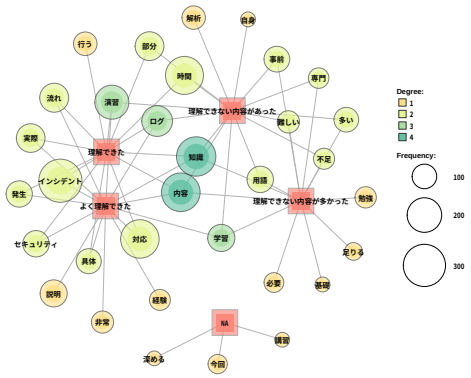

図 5　外部変数に変数「理解」を投入した共起ネットワーク図

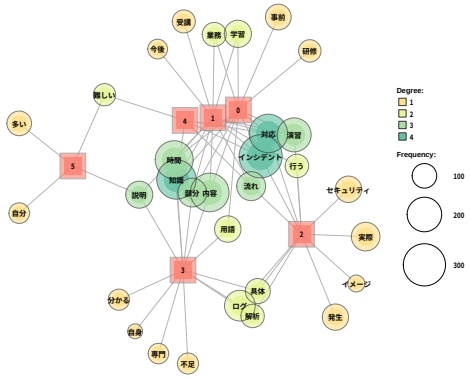

図 6　外部にクラスタ番号を投入した共起ネットワーク図

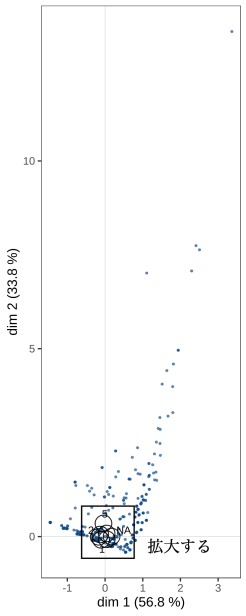

図 7　外部変数なしの共起ネットワーク図