

4-2 構造化データ解析で選択肢回答空間と機械学習による自由記述部分のクラスタリング結果を結合する

－ CYDER 受講者アンケートの分析手法の開発と今後の課題 －

4-2 *Projecting the Clustering Result of Free-Description Portion by Machine Learning to the Reference Space with Choice Answers Using Structured Data Analysis* - Development of Analytical Methods for CYDER Participant Questionnaires and Future Challenges -

藤本 一男 根本 颯汰

FUJIMOTO Kazuo and NEMOTO Sota

本稿が説明するのは、NICT サイバーセキュリティ研究所ナショナルトレーニングセンターが実施する実践的サイバー防御演習 CYDER の受講者アンケートの分析手法である。アンケートの選択肢回答部分には多重対応分析を用いて参照空間を生成し、自由記述回答には機械学習による非階層型クラスタリングを適用し、得られた要約を参照空間に射影した。生成した参照空間は、1) 受講者の満足度、2) 組織のセキュリティ対応、3) 受講者属性の3つである。クラスターは6つに分類され、それぞれに要約ラベルを付与した。構造化データ解析の手法でこのラベルを参照空間に射影し、回答選択肢の範囲を超えた分析が可能であることを確認した。全体的に受講満足度が高い中でも、インシデント対応に関する反応が確認された。無回答は、そこそこの満足をしている受講者に多く、不満の表明ではないことも確認された。

This paper describes an analysis method for participant questionnaires of CYDER, a practical cyber defense exercise by the National Training Center of NICT Cybersecurity Research Institute. Multiple Correspondence Analysis generated a reference space for the multiple-choice responses, and Non-Hierarchical Clustering by machine learning was applied to the free responses, projecting the resulting summary onto the reference space. The generated reference space consisted of 1) participant satisfaction, 2) organizational security measures, and 3) participant attributes. Clusters were classified into six (+ blank) and each assigned a summary label. These labels were projected into the reference space using Structured Data Analysis techniques, confirming the feasibility of analysis beyond the response options. Even among the high overall course satisfaction, responses related to incident response were confirmed. Non-responses were common among participants satisfied with the course, indicating they were not dissatisfied.

1 分析対象と問題の所在

1.1 分析対象

我々の研究の分析対象は、NICT サイバーセキュリティ研究所ナショナルトレーニングセンターが実施している実践的サイバー防御演習 CYDER^{*1} での受講者アンケートである。主要な集合演習のコースは、A、B (B-1、B-2)、C の3つがあり^{*2}、年間受講者数は約3,000名である。

この受講者アンケートの調査票は、選択肢回答部分と自由記述部分から構成されている。選択肢回答部分は、1) 回答者の個人属性に関する部分と 2) 所属組織

に関する部分に、そして、3) CYDER の中心プログラムである集合演習への評価部分とに分けることができる。

*1 CYDER (Cyber Defense Exercise with Recurrence : 実践的サイバー防御演習) とは、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が実施する、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である。https://cyder.nict.go.jp/

*2 A コースは初級。B コースは中級。B-1 コース: 地方公共団体勤務で、情報システム経験2年以上、B-2 コース: 国の機関など、重要社会基盤事業者、民間企業で、同じく情報システム経験2年以上。C コースは準上級。経験3～4年。これ以外にオンラインコースがある。詳細は、https://cyder.nict.go.jp/course/index.html を参照。

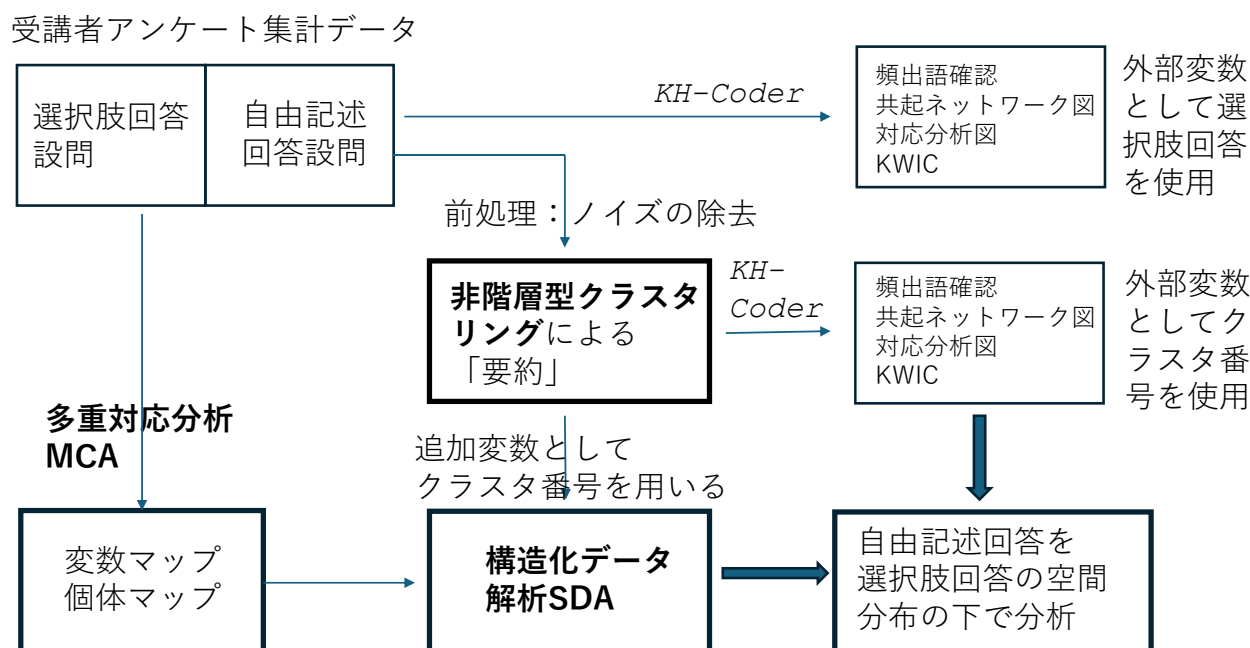


図 1 分析手法の構成図

この選択肢回答部分は、年齢を除いてカテゴリカル変数である。しかし、年齢も1歳刻みでは分析結果の解釈が困難であるので、10歳刻みの変数にリコードしている。そのため、全てがカテゴリカル変数である。

選択肢として用意できていない内容は、自由記述部分に記入されることを期待している。調査票は付録A-1を参照のこと。

1.2 問題の所在

この実践的サイバー防御演習 CYDER はサイバーセキュリティの事案が発生した際のインシデント・レスポンス(情報システム部門の担当者がインシデント発生時にとるべき行動)についての実践的集合演習であり、受講者の参加モチベーションは非常に高い。そのため、受講満足度に関係する変数においては、ほとんどが満足傾向の回答が記録されている。そうした傾向を確認するだけであれば、アンケート収集の意味がない。

サイバーセキュリティの領域では、日々新しい脅威が発生しており、そのため、回答選択肢の形では設問を構造化できない要素も多い。そうした点も踏まえて、受講者の声を吸い上げ、演習の質を向上させるためにも、構造化された選択肢回答と自由記述に現れた構造化されていない部分からの受講者の注目点、要望、集合演習の課題の「発見」が求められる。

そこで我々は、選択肢回答設問と自由記述回答部分を連携することで、構造化されていない受講生の声を抽出することを行う方法を検討した。

なお、本稿は、言語処理学会 30 回大会 (NL2024) で

根本・藤本が発表したもの [1] を基礎に発展させたものである。

2 分析手法の構成

2.1 二つの構成要素：多重対応分析と機械学習による非階層型クラスタリング

本稿が説明する分析手法の全体像を、図1に示した。

(1) 分析対象である集計データの「選択肢回答設問」部分に対して多重対応分析 (Multiple Correspondence Analysis: MCA) を行い変数空間と個体空間を生成し、(2) 自由記述回答に対しては、言語モデルを用いたアプローチによる非階層型クラスタリングでクラスタへの分割を行い、クラスタごとに要約ラベル付け、(3) 各クラスタを、選択肢回答との関連で分析するために、(1) で生成された空間に射影し、解釈を行う。

2.2 構造化データ解析：多重対応分析が生成した空間にクラスタ番号を射影する

この分析手法は、多重対応分析による参照空間の生成と、言語モデル・クラスタリングによる自由記述部分の要約(ラベル付け)の部分から成り、後者を前者に射影する。

この射影が可能になる原理は、多重対応分析において追加変数(Supplementary Variables)が、空間生成に影響することなく座標を取得できるということによる。この射影は、調査項目変数によって参照空間を生成し、そこに、空間生成には寄与させなかった人口統計学的変数(年齢、性別など)を射影し、空間の構造を

解釈する、と説明されている。こうした空間生成とそれに対する追加変数の射影による分析手法は、構造化データ解析 (Structured Data Analysis: SDA) として定式化されている [2][3]*3。

多重対応分析では、個体は何らかの統計量に抽象化されることなく、個体のまま座標をもったポイントとして参照されるため、選択肢回答設問に対する量的調査とインタビュー調査のような質的調査を結びつけることを可能にする。『多重対応分析』[3] 第1章の題辞の下には、「幾何学が量と質を架橋する」とあり、こうした混合研究法についての言明が記されている*4。

本稿で説明している自由記述部分に対するクラスタリングによる変数生成とその参照空間への射影は、こうした多重対応分析を用いた混合研究法の一つである*5。

2.3 自由記述回答部分の言語モデルによるクラスタリング = 要約

我々は、各回答者が回答した自由記述回答からクラスタリングによる要約によって、新たな変数として構成し、それを追加変数として、回答選択肢によって生成された空間に射影した。

ここで射影される空間を生成する変数が異なれば、つまり射影する空間が異なれば、クラスタ番号の空間内での位置は異なったものになる。参照空間生成に用いる変数として、I) 受講の満足度関連の変数、II) 受講者の所属組織のセキュリティ対策度関連の変数、III) 受講者の属性に関する変数、と3種類の空間を生成し、そこに自由記述部分のクラスタリング要約を射影し、配置を検討した。

こうして選択肢回答変数で生成された参照空間に、クラスタリングされた自由記述回答を射影することで手に入った幾何学的な配置の空間マップと KWIC コンコーダンスをもとに、自由記述部分を含んだ総合的な分析が可能になった。

なお、言葉による意味の生成は、それが発話される場との関係で解釈されるものと考えらるならば、このように射影空間を定義して解釈の前提とすることは重要である。

3 対応分析／多重対応分析とはどのような分析手法か

我々がカテゴリカル変数である選択肢回答の分析 = 空間生成に用いたのは、多重対応分析という手法である。

本節では、この分析手法の原理的な構成と、空間生成の原理、そして、その空間への追加変数の射影の仕

組みを簡単に説明する。対応分析／多重対応分析の数理的な詳細は [5] を、構造化データ解析の方法的詳細は [2][3] を参照のこと。

3.1 対応分析と多重対応分析

対応分析 (Correspondence Analysis: CA)、多重対応分析は、数理的には、林知己夫の数量化手法 III 類と等価であると語られる手法である。つまり、数値ではないカテゴリカル・データをある基準の下で数値を付与する手法である。

この手法は、対応分析では2元クロス集計表を、多重対応分析では個体×変数のような調査集計表を処理対象とする。前者は2変数、後者は多変数だが、この多変数は、変数ごとの選択カテゴリに分解、展開 (Indicator 行列化) され、選択されているカテゴリに1をたてる、という形式に変換し (Crisp 行列化、Complete disjunctive 行列化)、その2元表に対する対応分析として処理される (古典的 MCA の一つ)。そのため、どちらの場合でも、処理対象は2元表である。

数理的にはこの2元表の各セルの期待値からの残差 (残差行列) に対して特異値分解 (Singular Value Decomposition: SVD) を用いて特異値を含めた3行列の積に展開し、必要に応じて次元縮減を行う。こうして、行変数と列変数を図示するのに必要になる座標が取得される。多重対応分析の場合は、多変数の変数カテゴリ同士の関係、個体の分布状況が図示される。

このように、生成された空間に座標を持つことが対応分析による「数量化」である。対応分析では、分析の単位は、行プロファイル・ベクトル、列プロファイル・ベクトルである。個体に対応している行のベクトルは、その個体がどのカテゴリを選択したのかが反映されている。つまり、分析単位は、回答パターンである。そのため、対応分析のフランスでの始祖であるベンゼクリも日本の林知己夫も、多変量解析という表現よりは、多次元データ解析という表現を好んでいると Greenacre は日本語版への序でコメントしている [5]。

*3 空間を生成させる変数を、アクティブ変数と呼び、その空間に射影する変数を追加変数と呼ぶ。このように変数を区別することが、構造化モデリングと呼ばれる [3]。

*4 こうした構造化データ解析の適用例とし、選択肢回答による量的調査と質的調査であるインタビュー調査を組み合わせる研究も行われている [4]。

*5 言語処理学会 NLP2023 に藤本・大畑で発表した「多重対応分析とアスペクトベース感情分析を組み合わせた受講者満足度調査データの分析手法の開発」[7] も考え方は同じである。ここでは、自由記述部分から抽出した重要頻出語句について、ABSA (アスペクトベース感情分析) という手法を用いて、抽出語にテキストの文脈からポジティブ／ネガティブのタグを付与し、それを数値化して選択肢回答空間に射影している。

MCA模試図的に…

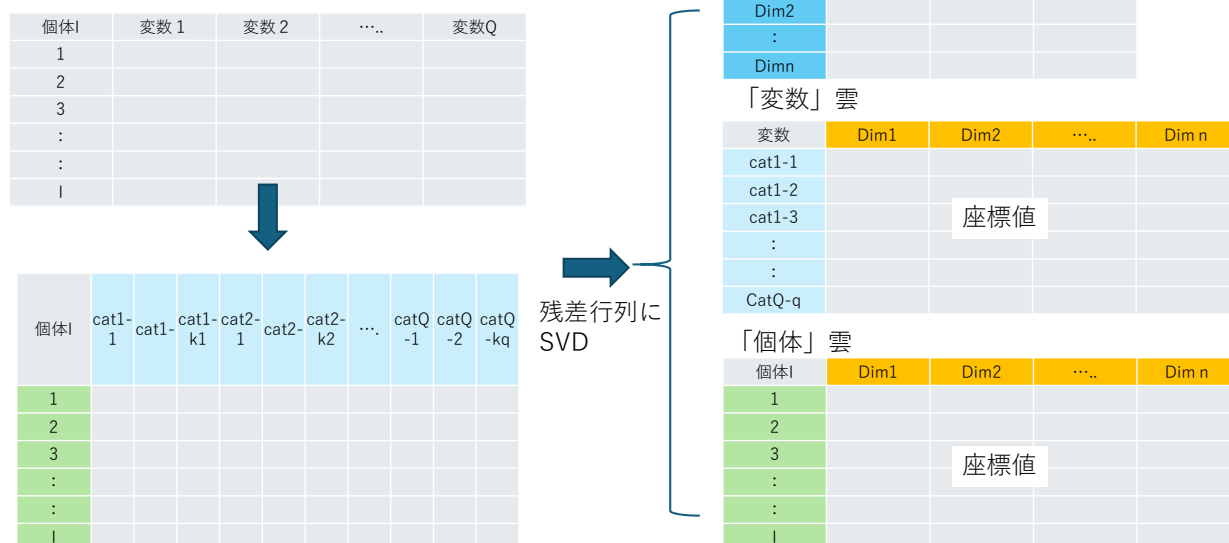


図2 変数(列)空間と個体(行)空間の生成

3.2 参照空間への射影を実現する追加変数をめぐる数理

対応分析は、先述のように行プロファイルと列プロファイルごとに空間を生成するのだが、そこではプロファイル間の距離を χ^2 距離で評価する。そのことが、似た分析手法である主成分分析 (Principal Component Analysis: PCA) では実現されない対応分析の特性を実現する。

行プロファイル空間と列プロファイル空間は、空間としては別空間であるが、座標軸に体现される分散は同じものである。そのために、2つの空間を重ねてあわせて行変数と列変数(CA)の、また個体と変数(MCA)の分布傾向を比較分析することが可能になる。

また、行プロファイル点と列プロファイル点は、遷移公式 (transition formula) に表現される関係で相互に関連しており、一方 (例えば行空間) の点は他方 (列空間) のすべての点の総計になる、という関係にある。

空間生成に用いられたポイントは、質量 (周辺度数) を有しており、それが座標軸を「引き寄せている」[5]。そこで質量が0であるポイントを考えてみると、このポイントは、空間生成には寄与しないが、遷移公式によって空間に座標値をもつことができる。これが、追加変数を参照空間に射影する原理である。

4 自由記述回答に対するクラスターリング

今回我々は、自由記述回答の構造化を目的に設定したが、このような自由記述を含むテキストの自動分析は以前から様々なアプローチで検討されてきた。ここで概観しておく。

4.1 テキスト・マイニング

簡単かつオーソドックスな手法として、文章に対する単語の頻出度 (term frequency: TF) や文書単位に対する単語の頻出度 (inverse document frequency: IDF) などの統計的尺度を用いて分析するテキスト・マイニングと呼ばれるアプローチがある [6][8]。

具体的にはテキストを単語や文節などの単位で区切り、それらの出現の頻度や共出現の相関、共起関係、出現傾向、時系列などを解析することで有用な情報を取り出す。

KH Coder [6] もこのアプローチを自動化したアプリケーションの一種であり、これを用いることで簡単にテキストの特徴を分析し、可視化することが可能になる。

しかし、このアプローチで得られる統計量は単語や文節単位で分析されるため、文脈情報が考慮されず、得られる特徴には限りがある。また、これらの統計量の解釈は分析者の問題意識などに依存するため、分析結果には分析者の主観が含まれる可能性がある。

こうしたことは、客観性に関して考慮すべきことであり、テキストの分析をするにあたっての課題となっている*6。

4.2 言語モデルによるアプローチ

テキスト・マイニングの一方で、近年の機械学習分

*6 解の客観性については検討すべきことがある。KH Coder では第1フェーズでは、誰がやっても同じ結果が生成されるが、第2フェーズでは、研究者の主体的責任においてコーディングルール (概念化) を設定し、研究者に固有の分析を可能にしている。

野の発展によって言語モデル (language model: LM) を用いたテキスト分析の新たなアプローチとなっている [9][10]。

言語モデルとは Wikipedia などの大規模なテキストのデータセットをディープラーニングの技術を用いて人力では獲得できない潜在的なパターンを学習するプログラムである。

特に ChatGPT (Chat Generative Pre-train Transformers) などの大規模言語モデル (large language model: LLM) [14][15] の台頭によって、言語モデルの活用はここ数年でより身近になっている。

これらの言語モデルには大まかに Encoder モデル [9][10] と Decoder モデル [15] (双方を兼ね備えた Encoder-Decoder モデルも存在するが、ここでは省略する) に分けられる。

Encoder モデルは 2017 年に公開され、機械学習分野における言語処理分野の発展のきっかけとなった BERT (Bidirectional Encoder Representations from Transformers) [9] を代表とし、学習したテキストデータをもとに数百～数千次元の埋め込み表現 (ベクトル表現) を作成でき、この処理を Embedding と呼ぶ。

この埋め込み表現を用いて分類器の構築やカテゴリ化を行うことで、SNS の内容から感情を予測する感情分析タスクやニュースの記事からジャンルを分類する文書分類タスクなどの自然言語を用いた様々な分類タスクを高精度に行える。

一方の Decoder モデルは生成モデルとも呼ばれ、先の ChatGPT がその一種であり、テキストの指示 (プロンプト) に応じて、埋め込み表現ではなくテキストをそのまま出力できる。

プロンプトに対してそのままテキストが出力されるため、API などを駆使することで専門家でなくとも手軽に使用できることから現在特に注目を浴び、様々な応用がされている。

先に紹介した単語や文節単位で処理するテキスト・

マイニングと異なり、言語モデルによってテキスト全体の埋め込み表現の取得が可能になるため、文脈を考慮した分析が可能となった。また、これらの言語モデルは使用者に依存せず、自動的かつ専門的な知識を必要とせずに使用することも可能である。

これよりテキスト・マイニングで課題となっていた、文脈情報の考慮や分析者の偏見の包含は一定量克服され、高精度かつ解釈可能なテキストの分析が可能になった。

4.3 非階層クラスタリング

そこで我々は言語モデルを用いたアプローチを選択し、その言語モデルによって得られた埋め込み表現を可視化、定性的な分析をするために、非階層型クラスタリングを行った。

非階層型クラスタリングとは言語モデルによって得られた各テキストの埋め込み表現から空間を生成し、各埋め込み表現の距離に基づいてあらかじめ決めたクラスタ数に合わせて各点を自動で分類する手法である [11][12]。

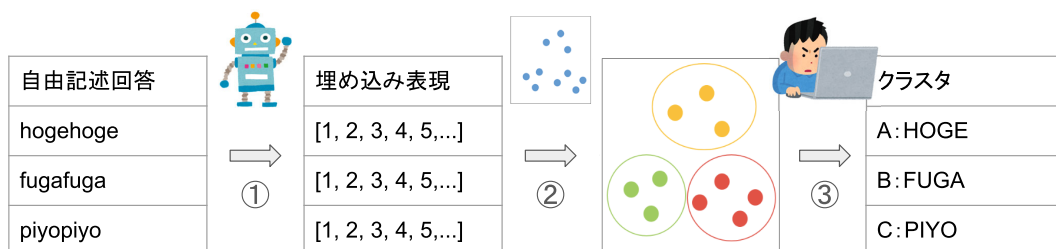
具体的には、各クラスタの重心が最も離れるように設定する K-means 法 [11] を拡張し、クラスタの重心の距離を考慮して自動でクラスタ数を設定する X-means [12] と呼ばれる手法を用いてクラスタリングを行った。

つまりこの手法を用いて自由記述内容が近い回答同士をクラスタとしてまとめることで、すべての自由記述回答を見る必要なく、クラスタを確認することで大まかに回答傾向を得ることができる。

クラスタを生成した後に、クラスタを構成する回答を TF-IDF などの統計的尺度を参考に分析者がタイトルをつけることで自由記述回答全体の要約を行った。

以下に今回我々が自由記述回答に対して行ったアプローチの概略図を示す。

我々のアプローチは 1) 自由記述回答を BERT の Embedding によって埋め込み表現に変し、2) 取得した



- ①: 言語モデルによる自由記述回答の埋め込み表現の取得
- ②: 埋め込み表現空間の作成とクラスタリング
- ③: 各クラスタの命名 (要約)

図3 自由記述回答に対する処理ステップ

埋め込み表現による非階層型クラスタリング、3) クラスタに対応した回答群を基に分析者がクラスタのラベル(要約)を生成する、という3ステップから成る。

このアプローチを自由記述項目である「気づいたことは何ですか?」(アンケート 20)、「理解できた or 理解できなかったと思う理由は何ですか?」(アンケート 15 - 2)「全体の感想について教えてください」(アンケート 22)の3項目に対して実施した。

5.2に「理解できた or 理解できなかったと思う理由は何ですか?」(アンケート 15 - 2)という問いに対する回答を上記アプローチで分析した結果を示す。

5 構造化データ解析による回答選択肢空間への自由記述の射影

5.1 基本集計の確認

実際の分析過程では、選択肢回答部分については、基本集計として、変数ごとの度数分布表、また、分布を確認するための棒グラフを作成した。変数ごとの基本集計に加えて、クロス集計などの組み合わせ、また年次比較、コース比較を行う可能性があるため、Shiny

Server^{*7}を用いて、分析サポートサイトを立ち上げた。

自由記述部分に対する基本集計に相当するものは、形態素解析に基づく語レベルでの集計である。我々はこの作業に KH Coder [6]^{*8}を用いた。KH Coder には、KWIC コンコーダンス表示機能があるので、総合的な分析過程でも活用している。

5.2 「理解」回答の項目を選択した理由についての自由記述のクラスタリング

4で紹介したアプローチによって得られたクラスタのラベルを表1示す。

この結果から BERT を用いたクラスタリングとその要約によって自由記述回答の全体像の大まかな内容理解と構造化ができた。

要約例においても実際の回答例と照らし合わせてみて、クラスタリングを行うことでクラスタを代表するような要約を簡単に行うことができた。

抽出された要約は以下の6項目である。これに「回答なし」(無記入)を加えて7カテゴリの変数とした。

以降ではこのクラスタリングの結果を用いて選択肢回答空間への射影を試みた。

5.3 三つの参照空間と「理解理由」の自由記述の分布

先に述べたように、自由記述部分を射影するものとして三つの参照空間を用意した。それは、「受講者満足度に関する空間」「所属組織のセキュリティ対応体制に関する空間」そして「受講者の個人属性に関する空間」である。

まず、自由記述部分に回答した(ON)か、否(OFF)かを変数化して射影し、回答あり/無しの分布を確認した。その結果、ON と OFF の分布の差が有意であるのが、「受講者満足度空間」だけであったので、以降の分析においては、この空間を使用することにする。所属組織のセキュリティ対応体制」「個人属性」は、このON と OFF に関して、有意な差異は見いだせなかった^{*9}。

なお、三つの参照空間と、三つの自由記述部分があるので、射影する組み合わせは9通りある。現在、そのうちの1つしか解釈対象にしてないが、この処理の自動化が可能になれば、この組み合わせのなかから有意な傾向を抽出できる可能性が残っている。これは、今後の課題にしたい。

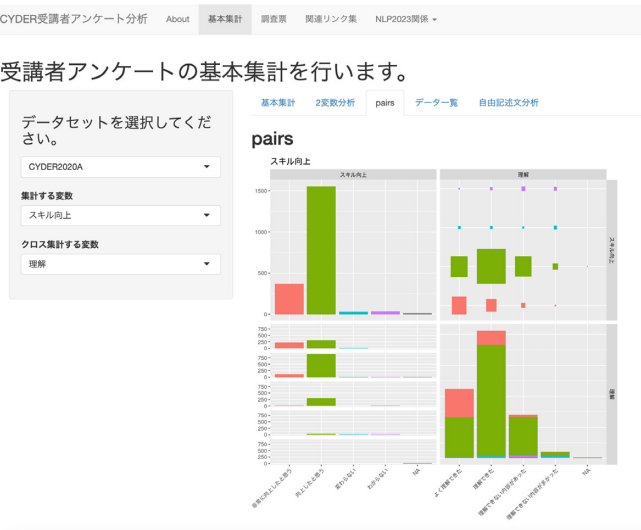


図4 Shiny Server で構築した基本集計ツール

表1 抽出されたクラスタと要約ラベル

クラスタ番号	要約ラベル
0	実務や研修の内容
1	インシデント対応経験と今後の方針
2	インシデント発生とその対応
3	講習に対する事前知識
4	講習の流れ
5	講師の説明

*7 <https://shiny.posit.co/>

*8 <https://kncoder.net/>

*9 検討は、相関比 η^2 で行っている。満足度関連空間、セキュリティ対応体制空間、個人属性空間ごとに、dim1 での値が、0.0108、0.00025、0.003730、であり、典型性検定の結果の p 値は、それぞれ、 $3e-6$ (≈ 0)、0.480、0.006 であった。

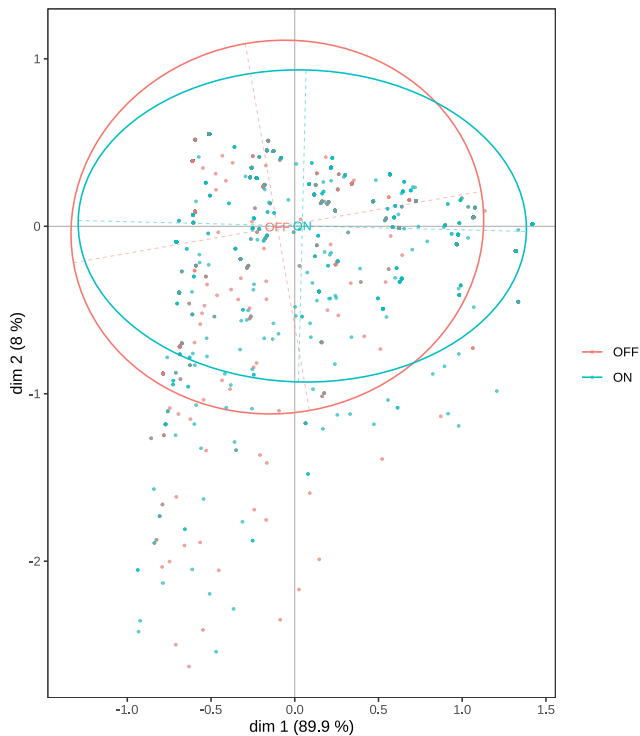


図5 満足度空間に理解度選択理由の自由記述あり ON／なし OFF の分散

これらを踏まえて、満足度関連変数によって生成された参照空間へのクラスタリングされた自由記述の射影を検討することにする。

5.4 満足度関連の変数によって生成された参照空間

満足関連変数としては、理解（アンケート 15 - 1）、講師説明（アンケート 16）、サポート適切（アンケート 18）、対応適切（アンケート 19）、進捗適切（アンケート 17）を選択し、これらとスキル向上（アンケート 14）変数で空間を生成した。

5.4.1 MOSAIC Plot^{*10} による単変数比較

MCA を行うまえに、クラスタリング要約ラベルと変数「理解」（アンケート 15 - 1）のクロス集計を行い、変数同士の連関度を MOSAIC plot で確認した（図 6）。このグラフからわかることは以下のとおりである。

- 1 行目の（よく理解できた）は、「2 インシデント発生とその対応」が、有意にプラスに多い。これは、

* 10 <https://cran.r-project.org/web/packages/vcd/vignettes/strucplot.pdf>,
https://www.researchgate.net/publication/2533471_A_Brief_History_of_the_Mosaic_Display,
<http://datavis.ca/mosaics/mosaics.pdf>

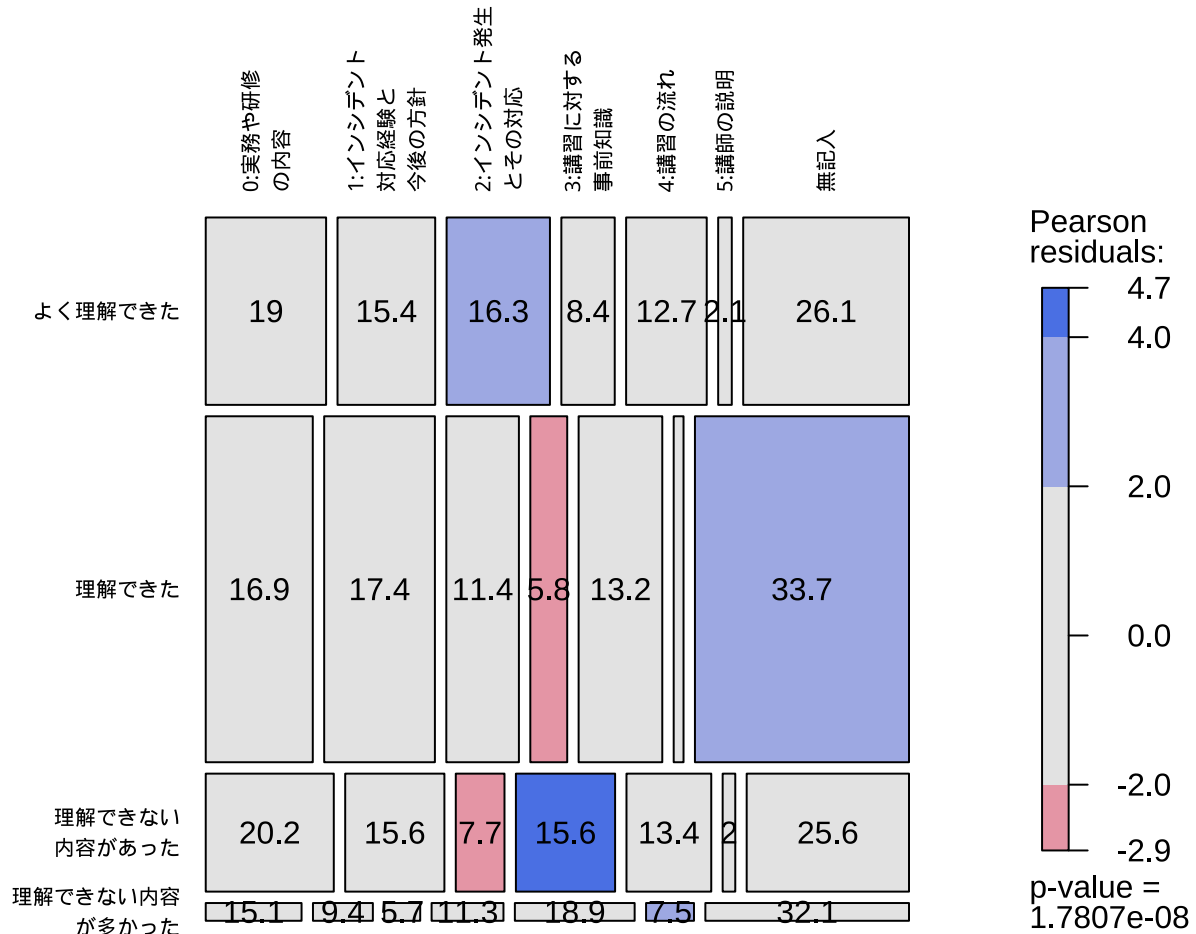


図6 変数「理解」15 - 1 と自由記述「理解理由」15 - 2 の要約の関係（数値は行％）

4 サイバーセキュリティに関する人材育成

講習の内容と「理解感」「手応え」が対応しているということが反映していると解釈できる。

- 「理解できた」の人の無記入が一番多い(期待値からの偏差が大きい)。「理解できた」ということは「可もなく不可もなく」ということで、特に書くこともなし、という反応の反映だと推察される。
- 同じく「理解できた」では、「2: インシデント発生とその対応」が、ピンクになっているので、期待値よりもマイナス側に偏位している。これも、受講者としては「特筆することなし」という反応だと考えられる。
- 「理解できない内容があった」という行(3行目)を見る。この「インシデント発生とその対応」のところがピンクであるが、対照的に「3: 講習に対する事前知識」が濃い青になっている。これは期待値(全体の平均)からみて、大きくプラスに偏位していることを示している。インシデント・レスポンスについては、特に書くこともないが、「理解できない内容があった」ということ理由を、「講習に対する事前知識の欠如」に求めていると考えられる。

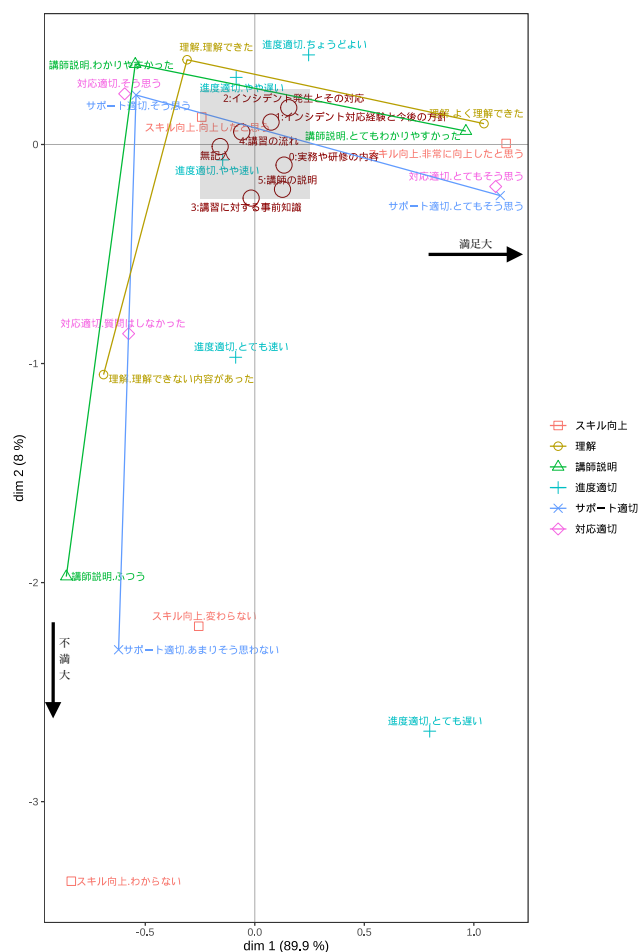


図7 「受講満足度」空間にクラスターリングによる「要約」を射影

- 4行目の「理解できない内容が多かった」は、「5: 講師の説明」に原因を求めている。

こうしてみると、クラスターリングによる分類要約とそれに付与されたラベルは、解釈可能であり、妥当性をもっているといえる。

5.4.2 この空間に抽出されたクラスターを射影する

受講者満足度空間のなかでの、このクラスターの参照空間での位置は、図7のようになっている(原点は全体の平均位置である)。この生成空間の軸の命名は、選択された変数のカテゴリの軸への寄与率から評価される。こうして命名された軸は、以下のようになる。

軸1右(+): 満足度大

軸2下(-): 不満度大

これを基に、図8の傾向をみれば、1軸右側に、無記入、4講習の流れについてのコメントが位置していることがわかる。また、2軸の下方向には、3講習の事前知識に関するもの、5講師の説明、0実務や研修の内容、が位置している。

このように、自由記述回答の傾向は理解度、満足関係の空間の解釈に整合していることが確認できた。

図8の第1象限に、インシデントに関するもの(1インシデント対応経験と今後の方針、2インシデント発生とその対応)が位置している。これは、講習の目的に沿って参加の意義が確認されていることを示している。

第2象限は、1軸の左、2軸の上という位置であり、「可もなく不可もなく」という「満足」の象限である。そこで描かれているものが、「4講習の流れ」についての

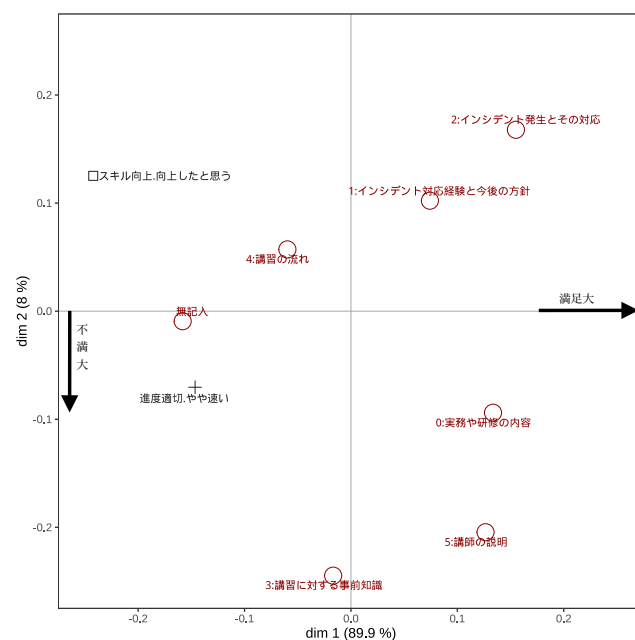


図8 図7のクラスター番号部分の拡大図

ものであり、あわせて(書くまでもない、よかったです!の結果と思われる)「無記入」が位置している。

第3象限は、下方向に「不満」が位置する。ここに「3 講習に対する事前知識」が位置しており、要求される事前知識への「不満」もしくは「反省」が位置している。

第4象限は、不満傾向と満足傾向が交差する象限だが、そこに、「0 実務や研修の内容」、「5 講師の説明」についての記述が位置している。

そもそも、満足度空間の軸1の形成に寄与しているカテゴリを確認すると「サポートへの適切度」(アンケート18「チューターのサポートは適切でしたか」)である。そして、軸2のマイナス方向の不満傾向への寄与は「講師の説明」(アンケート16「講師の説明はわかりやすかったですか」)カテゴリであった。

6 この手法によって得られた知見と課題

6.1 得られた知見

以上、2020年度に実施されたCYDER Aコースのデータをもとに、設問「理解」(アンケート15-1)の回答を選択した理由の自由記述(アンケート15-2)に対するクラスタリングによる要約ラベルを、受講者満足度関連変数で生成した参照空間に射影し、その解釈を行ってきた。

異なる参照空間に射影することで、当然のことながら、クラスタの位置は変化する。これは、機械学習によって「要約」された自由記述文がもつ多様性ゆえであり、どの文脈で解釈するかによって意味が異なることを表している。

今回の分析によって、CYDER受講者の自由記述に対する態度(記述するか無記入か)も含めて分析する際に手がかりとなるものを確認できた。さらに、集合演習終了後の限られた時間での受講者アンケートへの記入という時間的に厳しいデータ採取条件のもとで、入手できた受講者アンケートがもたらす受講者の反応を引き出すことができたと考えている。

6.2 今後の課題

クラスタリングによる自由記述へのラベルを変数として受講者満足度空間に射影したが、現状では、～に関する記述、という範囲にとどまっており、その～に関して「肯定的／積極的」もしくは「否定的／消極的」という分類ができれば、更に解釈の可視化が可能になる。大畑・藤本 2023 [7]では、自由記述文で用いられている頻出重要語句に対して、ABSA(アスペクトベース感情解析)を用いてポジティブ、ネガティブのタグをつけることを試みたが、この方向で文、パラグラフにタグをつける可能性を検討していきたい。

今回行った自由記述部分への機械学習による処理では、文字通りの意見文を対象にしている。しかし、今回分析対象にした調査票には、意見文以外にも、業務内容についての記述、所持している資格についての記述の部分がある。これらは、回答選択肢には該当するものがない場合に「その他」を選ばせ、その他の方は以下に記入してください、というように入力をうながす際に用いられる形式である。しかしその結果は、記入した人数分の別々の記述が並ぶことになる。こうした表現は記述文であっても、構造化し得るものからカテゴリを抽出することができれば、抽出を繰り返すことで、選択肢回答への構造化ができる可能性がある。こうした処理も検討していきたい。

多重対応分析による個体空間マップでは、個体は平均値や分散などの何らかの統計量に抽象化されことなく、個体としての位置(個体番号)を保持し続け、個体マップ上の特徴的な位置にある個体の回答、属性を確認することが可能である。個人情報を探査することに意味はないが、受講生の傾向を確認する目的で適用することは可能である。

今回の分析に使用したのは、2020年度に実施されたCYDER Aコースのものである。データは、2023まで保存されているので、回答傾向の年次比較、コース比較へと発展させていくことも可能である。分析フレームワークの確定を踏まえて発展させていく。

実際の分析過程については、多重対応分析／構造化データ解析に関する部分は、RとR用に開発されているパッケージを使用すれば、誰でも実施が可能であるが、機械学習の部分は、処理に関する専門知識を有したものが実施するしかない。その部分を補うために、テキストマイニングにおいてKH Coderが提供しているようなUIを開発する必要性を感じている。

謝辞

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のCYDER事業の皆様には、アンケート内容の変遷過程の説明を含めて多くのサポートいただきました。記して感謝いたします。

また、本稿で用いた、多重対応分析、構造化データ解析、幾何学的データ解析に関する知見は、科研費(20K02162)の助成を受けている「データの幾何学的配置に着目したカテゴリカルデータ分析手法の研究」の成果を用いています。

【参考文献】

- 1 根本 颯汰, 藤本 一男, “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,” 言語処理学会第30回

4 サイバーセキュリティに関する人材育成

- 年次大会発表論文集, pp.455-460, 2023.
https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/P2-1.pdf
- 2 Brigitte Le Roux and Henry Rouanet, "Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis," Dordrecht: Kluwer Academic Publishers, 2004.
 - 3 Brigitte Le Roux and Henry Rouanet, "Multiple Correspondence Analysis," SAGE publisher, 2010 (訳:大隅 昇, 小野 裕亮, 嶋真 紀子.『多重対応分析』オーム社,2021).
 - 4 Bennet, Tony, Michael Savage, Elizabeth Bortolaia Silva, Alan Warde, Modesto Gayo-Cal, and David Wright, "Cluture, Class, Distinction," 1972, (訳:磯 直樹, 香川 めい, 森田 次朗, 知念 渉, 相澤 真一.『文化・階級・卓越化』, 青弓社, 2017).
 - 5 Michael J, "Greenacre. Correspondence analysis in practice Third edition," Chapman & Hall/CRC interdisciplinary statistics series, CRC Press, Taylor & Francis Group, Boca Raton, 2017. (訳:藤本 一男『対応分析の理論と実践』, オーム社2020).
 - 6 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.
 - 7 大畑 和也, 藤本 一男, "多重対応分析とアスペクトベース感情分析を組み合わせた受講者満足度調査データの分析手法の開発," 言語処理学会第29回年次大会発表論文集, pp.255-260, 2023.
 - 8 Ramos, Juan, "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. vol.242, no.1, 2003.
 - 9 Devlin, Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.
 - 10 Liu, Yinhan, et al., "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692, 2019.
 - 11 Hartigan, John A., and Manchek A. Wong, "Algorithm AS 136: A k-means clustering algorithm." Journal of the royal statistical society. series c (applied statistics) 28.1, 1979: pp.100-108.
 - 12 Pelleg, Dan, and Andrew W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," Icml. vol.1, 2000.
 - 13 Touvron, Hugo, et al., "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
 - 14 Touvron, Hugo, et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.
 - 15 Raffel, Colin, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of machine learning research 21.140 pp.1-67, 2020.



藤本 一男 (ふじもと かずお)

ナショナルサイバートレーニングセンター
サイバートレーニング研究室
招へい専門員
津田塾大学
数学計算機科学研究所
特任研究員
対応分析、カテゴリカルデータ分析、社会調査法、コミュニケーション論

【受賞歴】

2024 年 言語処理学会 スポンサー賞



根本 颯汰 (ねもと そうた)

ナショナルサイバートレーニングセンター
サイバートレーニング研究室
リサーチアシスタント
法政大学大学院理工学研究科
応用情報工学専攻
博士課程
機械学習、深層学習、人工知能、自然言語処理
【受賞歴】
2024 年 言語処理学会 スポンサー賞

付録 A-1 CYDER 2020 年度 A コース 受講者アンケート 調査票

実践的サイバー防御演習



■ CYDER 受講後アンケートのご協力をお願い

(2020 年 7 月 1 日改定)

1. 本アンケートで収集する情報とその範囲

本アンケートでは、実践的サイバー防御演習 CYDER（以下「本演習」といいます。）の受講者の方から、アンケート項目記載のとおり、受講者の属性情報（性別、年齢、年齢、保有資格ないしスキル、本演習受講履歴、職種及び役職等）、受講者及びその所属組織における情報セキュリティへの取組みに関する情報、本演習に関する受講者の評価に関する情報等を収集させていただきます。

2. 収集する情報の利用目的等

1 で収集した情報は、国立研究開発法人情報通信研究機構ナショナルサイバーストレージングセンター（以下「当センター」といいます。）が、本演習の事業推進（当機構の演習・イベント等に関する情報提供、データ分析、取材、各種調査）のために利用するほか、演習品質の向上や演習環境改善等に関する研究開発のために利用します。また、当センターの業務補助業者が、演習運営支援に係る委託業務を遂行する限りで、1 で収集した情報を取り扱うことがあります。

3. 利用及び提供の制限

当センターでは、法令に基づく場合、不正アクセス、脅迫等の違法行為があった場合その他特別の理由のある場合を除き、1 で収集した情報を 2 の利用目的以外の目的のために自ら利用し、又は第三者に提供いたしません。ただし、統計的に処理された情報については公表することがあります。また、研究開発の成果として、学術論文等で発表をする際には、適切に個人の特定が不可能となるよう加工、又は、統計的なデータとして加工します。

4. ご協力をお願い

本アンケートは、受講に伴って強制されるものではなく、任意での協力をお願いするものですが、我が国におけるサイバーセキュリティ人材育成分野の発展のため、上記 1 ～ 3 に同意いただいた上、可能な限りご回答いただけますようお願いいたします。

実践的サイバー防御演習

令和 2 年度 実践的サイバー防御演習（CYDER）
アンケート

日付 _____ 受講者 ID _____ 氏名 _____

アンケート1 年齢をお答えください。
(_____ 歳)

アンケート2 性別をお答えください。
☐ 男 ☐ 女 ☐ その他 ☐ 無回答

アンケート3 あなたの職業をお答えください。
☐ 中央官庁公務員 ☐ 都道府県公務員 ☐ 市長村公務員 ☐ 経営者・役員 ☐ 会社員 ☐ 自営業・自由業
☐ その他

アンケート4-1 保有する情報系資格をお答えください（複数選択可）。
☐ CISM ☐ CISA ☐ CEH ☐ GSEC ☐ CISSP ☐ CompTIA Security+ ☐ CCENT
☐ ITパスポート ☐ 基本情報技術者 ☐ 応用情報技術者 ☐ 情報処理安全確保支援士
☐ 情報セキュリティマネジメント ☐ その他（アンケート4-2にもお答えください）

アンケート4-2 アンケート4-1で「その他」を選択された方にお伺いします。保有する情報系資格をご記入ください。

アンケート5 所属（出向元）組織全体の社員規模をお答えください。
☐ 50 名未満 ☐ 50 名～100 名未満 ☐ 100 名～500 名未満 ☐ 500 名～1000 名未満
☐ 1000 名～3000 名未満 ☐ 3000 名～10000 名未満

アンケート6 あなたの役職をお答えください。
☐ 経営者 ☐ 役員クラス ☐ 本部長クラス ☐ 部長クラス ☐ 課長クラス ☐ 主任／係長クラス
☐ 一般社員 ☐ 派遣社員 ☐ 契約社員 ☐ その他

アンケート7 あなたの日常業務をお答えください。

アンケート8 アンケート7で回答いただいた業務には、どの位の期間、携わっていますか？
☐ 1 年未満 ☐ 1～3 年 ☐ 4～5 年 ☐ 6～10 年 ☐ 11～20 年 ☐ 21 年以上

アンケート9 あなたの組織内で CSIRT にかかわる方は、何名位いらっしゃいますか？（ご本人や兼任者も含む）
☐ 5 名以下 ☐ 6～10 名 ☐ 11～20 名 ☐ 21～30 名 ☐ 31 名以上

実践的サイバー防御演習



アンケート10 あなたの組織内の CSIRT には、セキュリティに関する専門知識を持った方はいますか？

- ☐ 5 名以下 ☐ 6～10 名 ☐ 11～20 名 ☐ 21～30 名 ☐ 31 名以上
☐ 0 名、インシデント対応等、外部に委託している
☐ 0 名、インシデント対応等、外部に委託など特に対策をしていない

アンケート11-1 あなたの組織でインシデントが発生したことはありますか？

☐ はい (アンケート11-2 にもお答えください)

☐ いいえ

☐ わからない

☐ 無回答

アンケート11-2 アンケート11-1 で「はい」を選択された方にお伺いします、発生したインシデントについて可能な範囲で構いませんのでご記入ください。

アンケート12-1 セキュリティインシデントを想定した訓練・演習（講義も含める）を行っていますか？

- ☐ 定期的な訓練・演習を行っている (アンケート12-2 にもお答えください)
☐ 訓練・演習を行ったことがある (アンケート12-2 にもお答えください)
☐ 今後、訓練・演習を行う予定である (アンケート12-2 にもお答えください)
☐ 訓練・演習を行う予定はない
☐ わからない
☐ その他 (アンケート12-2 にもお答えください)

アンケート12-2 アンケート12-1 で「定期的な訓練・演習を行っている」「訓練・演習を行ったことがある」「今後、訓練・演習を行う予定である」「その他」を選んだ方にお伺いします。訓練・演習の頻度および一回あたりの時間、また、可能な範囲で構いませんので、訓練・演習の内容について、具体的に記入ください。

アンケート13 あなたの CYDER への参加回数をお選びください。

- ☐ 初めて参加した ☐ 2 回目 ☐ 3 回目 ☐ 4 回目 ☐ 5 回目 ☐ 6 回以上

アンケート14 今回の演習を受講して、受講前と比べて、ご自身のサイバーセキュリティに関する知識やスキルが向上したと思いますか？

- ☐ 非常に向上したと思う ☐ 向上したと思う ☐ 変わらない ☐ わからない

アンケート15-1 今回の演習は、理解できましたか？

- ☐ よく理解できた ☐ 理解できた ☐ 理解できない内容があった ☐ 理解できない内容が多かった

実践的サイバー防御演習



アンケート15-2 アンケート15-1 の回答を選択した理由について具体的に記入ください。ご意見・ご要望等をご記入いただいても結構です。

アンケート16 今回の演習で、講師の説明はわかりやすかったですか？

- ☐ とてもわかりやすかった ☐ わかりやすかった ☐ ふつう ☐ わかりづらかった ☐ とてもわかりづらかった

アンケート17 演習の全体的な進度は適切でしたか？

- ☐ とても速い ☐ やや速い ☐ ちょうどよい ☐ やや遅い ☐ とても遅い

アンケート18 今回の演習で、チューターのサポートは適切でしたか？

- ☐ とてもそう思う ☐ そう思う ☐ あまりそう思わない ☐ まったくそう思わない

アンケート19 今回の演習で、質問に対する講師、チューターの対応は適切でしたか？

- ☐ とてもそう思う ☐ そう思う ☐ あまりそう思わない ☐ まったくそう思わない ☐ 質問はしなかった

アンケート20 今回の演習の内容、進度、教材、講師、チューター、演習に用いた端末・CYDERANGE を含めた演習システム全体について、また、「これは気づきになった点」など良かった点や、改善を望む点がございましたら、ご記入ください。

アンケート21 来年度も CYDER を受講したいと思いますか？

- ☐ ぜひ受講したい ☐ 受講したい ☐ 受講の必要はない ☐ 受講したくない ☐ わからない

アンケート22 その他、演習について感想やご意見等ございましたら、ご記入ください。

アンケートにご協力いただき、ありがとうございます。