# 若手研究者助成金 実績報告書

令和　７年　４月２１日

日 本 大 学 学 長　　殿


| | |
|---|---|
| 氏　　名： | 関　　弘翔 |
| 所属・資格： | 理工学部・助教 |
| 実施研究所： | 理工学部・理工学研究所 |


下記のとおり報告いたします。

１　研究課題

| |
|---|
| 和文表記（40 字以内）<br>解釈可能なクラスタリングに向けた言語と表形式医療データの接続 |
| 英文表記<br>Connecting Language and Tabular Medical Data for Interpretable Clustering |

２　研究期間

令和　６年度

３　研究目的

本研究の目的は，言語と表形式医療データを接続し，共通の特徴表現を獲得することで，医療データのクラスタリング結果を言語によって解釈可能にすることである。

2022 年 3 月に日本医師会生命倫理懇談会が会長諮問「医療 AI の加速度的な進展をふまえた生命倫理の問題」について取りまとめた。医療 AI の健全な発展を促すためにいくつかの提言がされており，その一つが「人間が理解し，判断の根拠を説明できる医療 AI であること」である。説明可能な AI（explainable AI, XAI）への注目が高まり，さまざまな研究がなされているが，その技術を十分に活用し，利益を享受できるのは専門知識を持つ技術者であることが多い。そのため，誰しもにとって明快な説明といえる，言語に基づく XAI に関連した研究への期待は大きい。

自然言語と他の情報表現（モダリティ）を接続する研究は，言語と画像を対象とした CLIP（Contrastive Language–Image Pre-training）が 2021 年 1 月に提案されて以降注目され，Stable Diffusion に代表される画像生成 AI など，さまざまな応用展開がなされている。また，ChatGPT として広く認知された GPT（Generative Pre-training Transformer）など，自然言語領域における事前学習（Pre-training），特に自己教師あり学習技術の成功を受けて研究が深化している。

言語と表形式医療データを接続できれば，疾患分類やクラスタリングなど，医療データを用いて行うさまざまな機械学習タスクに応用できるため，貢献できる分野は多岐に渡る。

４　研究概要

本研究は，構造化された表形式の医療データと，人間が使用する言語，いわゆる自然言語を接続する特徴抽出方法の確立に取り組んだものである。近年，医療ビッグデータとも言える，現実の医療を反映した実臨床現場から得られるリアルワールドデータの利活用が，臨床試験などの医薬品開発や臨床研究領域において期待されている。医療ビッグデータの活用の一つとして医療 AI への関心が高まるなかで，透明性，説明可能性を備えた責任ある AI の実現が国際的に注目されている。医療現場における説明可能性は，技術者だけでなく医療従事者や患者にとっても明快である必要があり，それを実現する形は言語による説明に他ならないと考えた。しかし，表形式医療データと言語を接続する AI 研究は，画像や音といった非構造化データほど進んでいない現状がある。

そこで本研究では，大規模な表形式医療データを基にした自己教師あり学習により，言語と表形式データを接続（橋渡し）するような特徴表現を獲得する手法の確立を目指した。図 1 に提案の概要を示す。

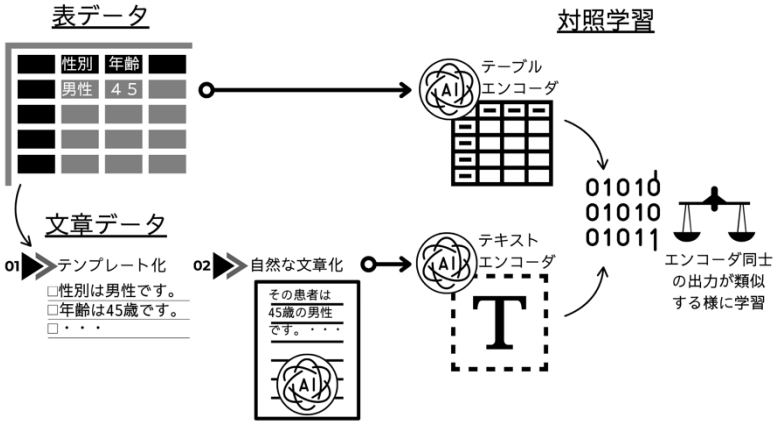



図 1　提案手法の概要図

まず，自然言語処理を応用して表形式データの分類を行う TabLLM[1]を参考に，表形式データの文章表現を得た。具体的には「属性名は xx である」というテンプレート（例えば「性別は男性である。年齢は 45 歳である。」など）を利用したテンプレート文章化を行った。さらにテンプレート文章を大規模言語モデル（Large Language Model, LLM）通すことによってより自然な文章化（例えば「その患者は 45 歳の男性です。」など）を行った。

文章化したデータを特徴ベクトル化するテキストエンコーダと，表形式データをそのまま入力として特徴ベクトル化するテーブルエンコーダを用意し，CLIP（Contrastive Language-Image Pre-training）[2]のように同一内容を表す文章と表のペアの特徴ベクトル同士が類似するように対照学習を行った。これにより，言語的な解釈の可能性を残した表データの特徴抽出器となるテーブルエンコーダを獲得できると考えた。

さらに，テーブルエンコーダによって得られた特徴ベクトルを k-means++法によりクラスタリングし，各クラスタの特徴分析を行うとともに，クラスタリング結果の言語による解釈を試みた。

[1] S. Hegselmann, et al. “TabLLM: Few-Shot Classification of Tabular Data with Large Language Models.” Proc of Int. Conf. on Artificial Intelligence and Statistics. PMLR, pp. 5549-5581, 2023.

[2] A. Radford, et al. “Learning Transferable Visual Models from Natural Language Supervision.” Proc. of Int. Conf. on Machine Learning. PMLR, pp. 8748-8763, 2021.

５　研究結果

本研究は日本大学薬学部臨床研究に関する倫理審査委員会の承認（20-012）を得て実施した。本研究で使用した RWD は，「慢性腎臓病」「認知症」「２型糖尿病」「肺がん」「リウマチ」「移植患者」の計６つの疾患等からなる総計約 300 万人規模の患者データである。これらの RWD は，メディカル・データ・ビジョン（株）より購入し，許諾を受けて使用した。

本研究では，薬物動態解析を行う研究や合成患者データを生成する研究への応用も見据え，使用する患者属性として関連研究で使用している，「性別（SEX）」「年齢（AGE）[year]」「体重（TBW）[kg]」「身長（HT）[cm]」，さらに臨床検査値として肝機能値の「ALT[U/L]」「AST[U/L]」および，腎機能値の「SCR[mg/dL]」とした。表１に使用した学習データの基本統計量を示す。なお，各疾患の件数は均等となるように調整している。

表 1. 学習データ（4800 件）の基本統計量

| | AGE[year] | TBW[kg] | HT[cm] | ALT[U/L] | AST[U/L] | SCR[mg/dL] |
|---|---|---|---|---|---|---|
| 平均 | 74.3 | 55.8 | 156.9 | 25.8 | 33.4 | 1.39 |
| 標準偏差 | 12.9 | 13.3 | 10.2 | 51.6 | 87.1 | 1.78 |
| 最小値 | 1.0 | 9.7 | 50.0 | 1.0 | 3.0 | 0.19 |
| 中央値 | 76.0 | 54.5 | 157.0 | 17.0 | 22.0 | 0.83 |
| 最大値 | 100.0 | 164.1 | 191.0 | 1866.0 | 3139.0 | 20.74 |

まず，表形式データのテンプレート文章化を行った。ここで使用したテンプレートは「The Sex is ${SEX}. The Age is ${AGE}. The Weight is ${TBW}. The Height is ${HT}. The AST level is ${AST}. The ALT level is ${ALT}. The Serum creatinine(SCR) level is ${SCR}.」のように作成した。${AGE}などには，実際の表データの値が代入される。${SEX}は男性の場合「male」が，女性の場合「female」が代入される様にした。SCR については LLM が解釈しやすいように正式名称を書いている。なお，疾患に関する情報は入力として用いないこととした。

テンプレート文章から自然な文章への変換には，様々な機械学習タスクをゼロショットで実現するために調整された LLM である T0-3B[3]を用いた。これにより，例えば「A male is 45 years old. A man is weighing 50kg and is 160 cm tall. ・・・」の様な文章が生成される。

テキストエンコーダには，医療用のデータセットで調整された LLM である ClinicalBERT[4]を採用した。テーブルエンコーダには，言語モデルで成功を収めた Transformer を表形式データに応用した先駆けである FT-Transformer[5]を採用した。ClinicalBERT が出力する特徴ベクトルは 768 次元であるが，表形式データに対しては高次元過ぎるため，全結合層を用いて 128 次元となるように射影した。FT-Transformer の１属性に対するトークン数は４とし，テキストエンコーダと同様に，全結合層を用いて出力が 128 次元となる様に射影した。それぞれの特徴ベクトルを L2 正規化した上で，同一データに対するテーブルエンコーダとテキストエンコーダの出力が類似する様に InfoNCE[6]で損失を計算し，Adam オプティマイザにより各エンコーダを最適化した。

学習したテーブルエンコーダを用いて出力した特徴ベクトルに対して k-means++法によりクラスタリングを行った。今回のデータは６つの疾患データから取得したものであるため，クラスタ数は試験的に６としている。また，128 次元の特徴ベクトルを次元削減手法の UMAP[7]により２次元化し，クラスタリング結果等を平面で可視化できる様にした。

〔５　研究結果（つづき）〕

図２に UMAP で次元削減した平面上で，クラスタごとに色分けした結果および，各クラスタの重心（図中の星）を可視化したもの示す。大きく塊が 2 つに分かれていることが確認できる。左の塊は 9 割以上のデータが男性であり，右の塊は 9 割以上のデータが女性であった。よって図より，各クラスタには性別が特定の偏りなく分布していることがわかる。

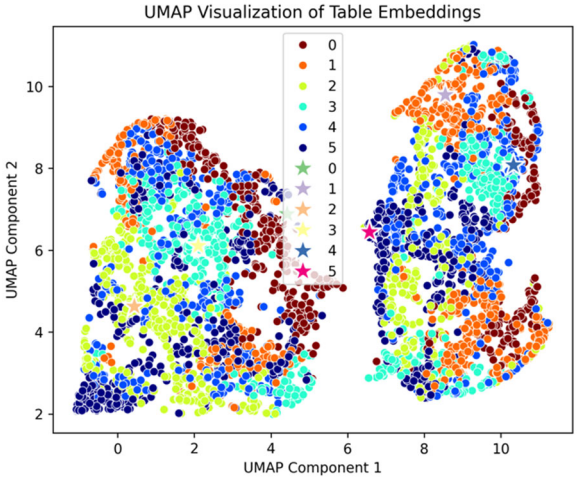



図２　クラスタリング結果の可視化

各クラスタのデータ数は 600~900 件で推移していた。ほとんどのクラスタで疾患は満遍なく分布していたが，クラスタ 5 は約 50%が慢性腎臓病の患者であり，特徴が見られた。腎機能マーカである SCR の平均は全体が 1.39 mg/dL に対してクラスタ 5 では 3.87 mg/dL であり，高い傾向が見られた。図 3 にクラスタごとの SCR のヒストグラムを示す。クラスタリング結果に対して言語による解釈が可能かを検証するため，テキストエンコーダに「This is a group of patients with abnormal kidney function values and high SCR level.」という文章を入力してベクトル化し，各クラスタの重心ベクトルとの類似度と比較したところ，他のクラスタに比べてクラスタ 5 との類似度が高くなった。また，「The patient had received an transplant」という文章をテキストエンコーダでベクトル化し，全学習データをテーブルエンコーダでベクトル化したものとの類似度を測り，上位 5 つのデータを確認したところ，4 つが実際に移植を受けた患者のものであった。よって，対照学習により構築したテーブルエンコーダの出力ベクトルは，言語によって解釈できる可能性が示唆された。

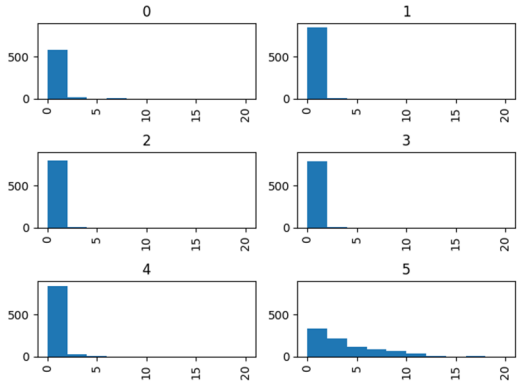



図３　各クラスタの SCR のヒストグラム

[3] V. Sanh, et al. “Multitask Prompted Training Enables Zero-Shot Task Generalization.” arXiv preprint arXiv:2110.08207, 2021.

[4] K. Huang, et al. “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.” arXiv preprint arXiv:1904.05342, 2019.

[5] Y. Gorishniy, et al. “Revisiting Deep Learning Models for Tabular Data.” Advances in neural information processing systems, 34: 18932-18943, 2021.

[6] A. Oord, et al. “Representation Learning with Contrastive Predictive Coding.” arXiv preprint arXiv:1807.03748, 2018.

[7] L. Mcinnes, et al. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” arXiv preprint arXiv:1802.03426, 2018.

６　将来獲得を目指す学内外競争的資金の種類

科学研究費助成事業基盤研究（C）•公益財団法人立石科学技術振興財団研究助成

７　本研究をもとに申請した外部研究資金及びその獲得状況

| 氏　　名 | 申請年度 | 研 究 種 目 | 研　究　課　題 | 採・否 |
|---|---|---|---|---|
| | | | | |

８　学会発表等

要旨集の抜き刷り，発表原稿のコピー等及び発表したことがわかるものを添付してください。

| 発表者名 | 学会，発表会名等 | 発表テーマ | 発表年月 |
|---|---|---|---|
| 宮野咲紀，関弘翔，川村駿太，青山隆彦，辻泰弘，細野裕行 | 第 45 回日本臨床薬理学会学術総会 | 限られたデータから臨床データ合成を創出する生成モデルの比較検証 | 2024 年 12 月 |

９　著書・雑誌論文

著書，雑誌，抜き刷り等を添付してください。コピーの場合は掲載されたことが客観的にわかるものを添付してください。

| 著者・執筆者 | 著書名・雑誌名／論文名 | 巻・号 | 年月 | 出版社・発行所 |
|---|---|---|---|---|
| | | | | |

１０　本研究課題に関するその他の発表（新聞，一般雑誌掲載等）

１１　他の研究への発展

本研究では，言語モデルと組み合わせることで，表形式データに対して言語的な意味合いも内包した特徴抽出方法を実現することを目指したものである。表形式データを文章化して扱うアプローチは，表形式データの分類，回帰や合成など様々な分野で注目されているが，言語モデルは大規模化しており，扱いにくい欠点がある。本研究で得られる表データの特徴抽出器は言語モデルと比較して軽量であるため，様々な機械学習タスクへの応用がしやすい利点がある。今後は，表形式データに対する回帰や合成の研究に，本研究成果を応用して発展させる予定である。

１２　その他の成果

１３　本研究課題による発明届の提出及び知的財産権の出願・取得状況

| 名　　称 | 知的財産の種類 | 発　明　者<br>（考案者・創作者） | 権　利　者 | 出願・登録番号 | 出願年月日 | 取得年月日 |
|---|---|---|---|---|---|---|
| なし | | | | | | |

**※ホームページ等での公開（☑可・□否）**いずれかをチェックしてください。
否の場合は，理由書を別途添付のこと。