類似データセット発見課題における詳細なデータセット分類に基づいた有効性の評価

作本猛 [1]　早矢仕晃章 [2]　坂地泰紀 [2]　野中尋史 [3,4]
[1] 長岡技術科学大学 [2] 東京大学
[3] 愛知工業大学 [4] 株式会社マヨラボ
s183353@stn.nagaokaut.ac.jp　{hayashi,sakaji}@sys.t.u-tokyo.ac.jp
hnonaka@aitech.ac.jp

概要

近年のコンピュータ技術の発展に伴い，互いに類似するデータセットを発見するための方法が盛んに研究されているが，従来研究ではデータセットを探索する個々のユーザのニーズを反映するような評価項目が用いられてこなかった．そこで，本研究では，類似データセット発見手法の評価において，データプラットフォームで実際に用いられている分類基準に基づいた 4 種類の詳細な項目を評価に適用した．その結果，変数ラベル間の Dice 係数を用いた手法が，課題やメイントピックが完全に一致するデータセットの探索に有用であることが示された．

1　背景

コンピュータ技術の発展とデータプラットフォームの台頭により，データの市場取引や Web 上でのデータ公開・共有が盛んになっている [1, 2]．また，機械学習・人工知能分野の技術発展によってこうした需要はより高まり，企業や行政組織においてもデータドリブンな分析や意思決定を重視する動きが増加している．市場に流通するデータセットが多様かつ膨大になる中で，データセット発見，特に，データセット間の類似性に基づく検索に注目が集まっている [3, 4, 5]．これは，ユーザの全てが求めるデータセットに関する知識を持っているという前提は現実的ではなく [6]，また，データセットのメタデータには品質のばらつきの問題があること [7] から，データセット検索において従来の Web 文書を対象とした検索技術が必ずしも適切とは言えないことが要因として挙げられる．また，Degbelo ら [8] が指摘するように，「自身の課題に関係しているが，想定・知識の外にある，まだ見ぬデータセットが存在するのか否かを知りたい」という重要なニーズの存在も類似性に基づくデータセット検索への注目を裏付けている．

こうした背景から，一部のプラットフォームでは，データセットの発見可能性を向上させるために独自のタグオントロジーを定義している．例えば，Kaggle では「新型コロナウィルス感染症」や「暗号通貨」といった，データセットのメイントピック的な要素を意味する subject や，テキスト，画像といったデータ型を意味する data type，データセットに関連する課題や収集目的を意味する task といったタグが存在する．データセットの分類として広く使用されている上記のような類似性の観点を評価に用いることで，類似データセットを探索する個々のユーザのニーズに適した発見手法を提示できると考えられる．しかしながら，従来の類似データセット発見の研究において，上記のような詳細なデータセット分類に基づいた評価には焦点が向けられてこなかった．

我々は，各手法によって得られたデータセットペア間に見られる類似性の種類や，具体的・抽象的といった類似性の段階を個別に評価するために，データプラットフォームで実際に用いられているデータセット分類を評価に適用した．その結果として，変数ラベル間の Dice 係数に基づく類似度が中〜高のデータセットペア間では，データセットに関連する学術的分野や課題については類似するが，それぞれの具体的なメイントピックは異なるといった，比較的抽象的な類似が多く見られることを確認した．また，類似度の数値に比例して，メイントピックについても類似したデータセットペアが増加するという傾向が見られたことから，変数ラベル間の Dice 係数に基づく類似度はデータセット間の内容的な類似

性を適切に反映していると考えられる．

本研究の主な貢献として，どのような種類の類似がデータセットペア間で見られやすいかを類似度の階級ごとに整理することによって，特定のデータセット探索のニーズに沿った手法選択に有用な知見を提供した点が挙げられる．

2 関連研究

類似データセット発見における評価は，データセット構造の類似性に焦点を当てたものと，データセットに関連するトピックの類似性に焦点を当てたものの 2 種類が存在する．データセット本体の構造を用いた評価の例として，ランダム分割されたデータセットにおける再現可能性を評価する方法 [9, 10] や，下流タスクの評価指標に基づいて評価を行う方法 [11, 12] などが挙げられる．これらの取り組みは，同じデータ形式や課題に関連するデータセットの集合，あるいは同じデータスキーマを持つテーブルの集合から，構造的に類似する部分集合を発見する性能の評価に焦点を当てており，より雑多なデータセットの集合を対象とする我々の取り組みとは相互補完的な立ち位置にある．

データセットに関連するトピックの類似性に基づいて評価を行った取り組みとしては，Sakaji ら [13] の取り組みが挙げられる．Sakaji ら [13] は内容的側面と地理的側面の 2 種類の項目に基づいて，各データセットペアがどの程度類似しているかを定量的に評価した．しかし，データセットの内容的な類似性を評価するための尺度は一つではない．例えば，Kaggle データプラットフォームでは，subject や task，data type のような，共通項目を持つデータセットへのアクセシビリティを高めるための数種類の分類が存在している．加えて，いくつかの分類には階層構造が定義されている．得られたデータセットペアがどの分類で類似しているのか，どの段階で類似しているのかについて定量評価が可能となれば，データセットを探索するユーザのニーズに合わせた手法の選択がより容易になると考えられる．そこで，我々の研究では，Kaggle や Papers with code といったプラットフォームで用いられているデータセット分類を適用することで，各データセットペアがどのような項目で類似しているか，具体的に類似しているか抽象的に類似しているかをそれぞれ定量的に評価している．

3 実験

図 1 に示すとおり，以下の手順に沿って実験を行う．(1) データセット集合から全ての可能なペアを作成する．(2) 全てのデータセットペアについて，後述する手法（3.2 節）に従ってデータセット間の類似度を計算する．(3) 類似度を 10 段階の階級（0〜0.1，...，0.9〜1.0）に区分し，各階級に対して対応する類似度を持つデータセットペアを割り当てる．(4) 各階級に含まれるデータセットペアから最大で 10 件のペアをランダムにサンプリングし，各評価項目（3.3 節）で共通するペアの割合（適合率）を計算する．

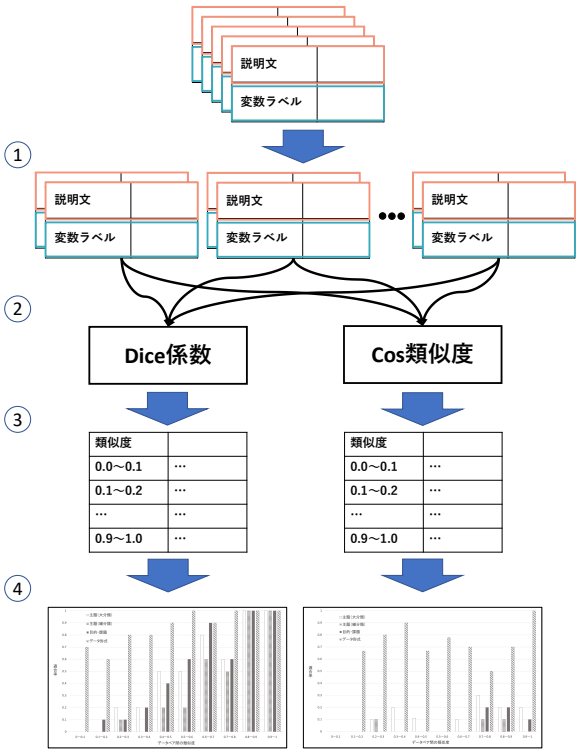

図 1 評価実験の概要

3.1 データセット

本研究ではデータプラットフォーム Kaggle[1] から 2020 年 1 月 20 日時点で収集されたデータセットのうち，少なくとも 1 件のタグを含み，かつ少なくとも 100 単語以上で構成される説明文を含んだデータセット 4041 件を対象に実験を行う．各データセットに含まれるメタデータは以下の表 1 に示すとおりである．

1） https://www.kaggle.com/

表 1　データセットに含まれるメタデータの概要

説明文	データセットの内容を詳述した文章
変数ラベル	データ属性の論理集合

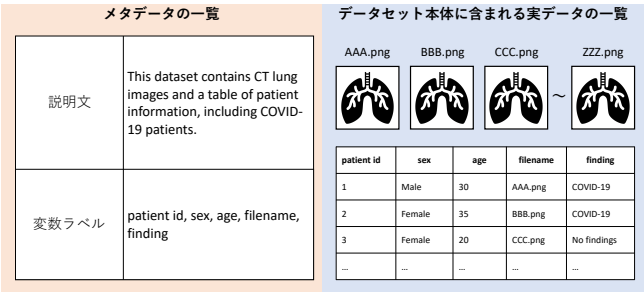

図 2　メタデータと実データとの対応関係

3.2　類似データセットペアの発見

今回の実験では，Sakaji ら [13] が提案した手法のうち，変数ラベル間の Dice 係数（以後，**変数**），説明文の BERT 埋め込みベクトル間のコサイン類似度（以後，**BERT**）の 2 種類の類似データセットペアの発見手法を用いる．

3.2.1　変数

変数を用いた類似度については，先行研究の方法 [13] に従い，Dice 係数を用いて以下のように計算する．

$$\mathrm{Dice}(d_i, d_j) = \frac{2|\mathrm{vars}(d_i) \cap \mathrm{vars}(d_j)|}{|\mathrm{vars}(d_i)| + |\mathrm{vars}(d_j)|}, \tag{1}$$

ここで，d_i，d_j はそれぞれ異なるデータセットを指し，vars(d) はデータセット d が保有する変数の集合を意味する．

3.2.2　BERT

データセットの内容が類似している場合，それぞれの説明文の埋め込みベクトル間のコサイン類似度は高くなると考えられる．そこで，先行研究 [13] で最も良い結果を示した，説明文の名詞を対象とした BERT 埋め込みに基づく手法を使用する．

$$T_W = \mathrm{BERT}(W), \tag{2}$$

ここで，$W = \{w_1, w_2, ..., w_N\}$ は長さ N の系列である．BERT(seq) は系列 seq を入力として受け取り，最終隠れ層のベクトル T_W を返す事前学習済みの BERT モデルである．最後に，入力系列長に関わらず比較可能なベクトル表現 $V_{BERT}(W)$ を得るために，以下のように平均化を行う．

$$V_{BERT}(W) = \frac{1}{|T_W|} \sum_{t \in T_W} t, \tag{3}$$

BERT の事前学習済みモデルには，英語版 Wikipedia で学習された bert-base-uncased[14][2] を使用する．

3.3　データセット分類に基づく評価項目

Kaggle では，データセットのメイントピックを意味する subject，収集目的や関連する課題を意味する task，データ形式を意味する data type といった分類が存在する．また，Papers with code には Modality や Task といった分類が存在しており，Modality は Kaggle における data type と同様の分類である．そこで，我々はこれらの分類を基に，以下の表に示す 4 種類の分類を定義し，各分類に基づいてより詳細な有効性の検証を行う．

表 2　データセット分類に基づく類似性評価項目

名称	具体例
主題（大分類）	医科学，経済，政治，ほか
主題（細分類）	Covid-19，株価，暗号通貨，ほか
目的・課題	感染者予測，価格変動，ほか
データ形式	テーブル，時系列，ほか

表 2 の 1，2 行目に示す主題（大分類），主題（細分類）は，いずれも Kaggle における subject と対応している．主題（細分類）はデータセットが主に対象とする事物（メイントピック）そのものを指しており，主題（大分類）はそれらが属する学術的，あるいは技術的な分野・領域を指している．表の 3 行目に示す目的・課題は，Kaggle の task や Papers with code の Task と対応する項目であり，主にそのデータセットを用いて解決したい課題を意味している．表の最下部に示すデータ形式は，Kaggle の data type や Papers with code の Modality と対応しており，テーブルや時系列といった項目が存在する．これらの項目に基づいたデータセット間類似性評価の一例として，特定の企業における株価の推移を日毎に記録したデータセットと，暗号通貨の価格変動を日毎に記録したデータセットの 2 種類を挙げる．両データセットは共に経済の分野に属しており，目的・課題やデータセットの形式も共通したものであると考えられるが，それぞれのメイントピックは株価と暗号通貨のように異なっている．これらのことから，両データセットは主題（大分類），目的・課題，データ形式において類似しているが，主題（細分類）については異なるとみなされる．

2）　https://huggingface.co/bert-base-uncased

4　実験結果

図 3,4 に，それぞれの手法 (変数，BERT) によって発見された類似データセットペアに対する，各類似度階級における各評価項目の適合率を示す．まず，どちらの結果においても，データ形式に関する適合率は全ての類似度の階級において高い値を示していることがわかるが，これは今回の実験で使用したデータセットの多くがテーブルデータセットであることに起因すると考えられる．図 3 より，変数に基づく類似データセットペアにおいて，類似度の数値に比例して，主題（大分類），主題（細分類），目的・課題の 3 項目の適合率が高くなる傾向が見られた．特に，類似度が 0.4〜0.6 の階級では，主題（大分類），目的・課題に対して，主題（細分類）の適合率が 2 分の 1 程度と低い値を示しているが，0.6 以上の階級ではこれらの差が縮小していき，0.8 以上の階級では全ての項目において適合率が最大の値を示している．これは，変数ラベルに基づくデータセット間類似度の数値は，データセット間の類似性の粒度の細かさ，言い換えると，データセット間で共通する項目の多さを適切に反映していると言える．以下の表 3 に，各階級において得られたデータセットペアの具体例を示す．

表 3　変数ラベル間の類似度に基づいて得られたデータセットペアの例．なお，主題（大），主題（細）はそれぞれ主題（大分類）と主題（細分類）を意味している．

階級	主題（大）	主題（細）	目的・課題
0.6〜0.7	**農産**	プラム	**生産量統計**
	農産	チェリー	**生産量統計**
0.7〜0.8	**経済**	暗号通貨	**価格変動**
	経済	S&P500	**価格変動**
0.8〜0.9	**医科学**	**Covid-19**	**感染者予測**
	医科学	**Covid-19**	**感染者予測**

対して，図 4 に示すように，BERT に基づく類似データセットペアは，データ形式以外の 3 項目における適合率が，全ての類似度階級において低い値を示した．類似度の高い領域 (0.7〜) では，主題（細分類）や目的・課題に沿った類似データセットのペアが増加するという傾向が見られたが，いずれも適合率が 0.1〜0.2 程度と低い値を示している．これは，説明文を持つ Kaggle データセットにおいて，テンプレート文章をそのまま説明文として流用したものが多く存在すること，また，教師なしのテキスト間意味的類似度（Semantic Textual Similarity）の測定において，BERT は必ずしも有効ではないこと [15] などが要因として考えられる．

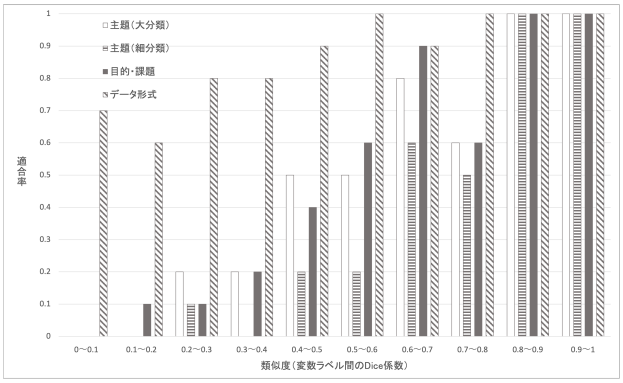

図 3　各評価項目におけるデータセットペア間の適合率（変数）

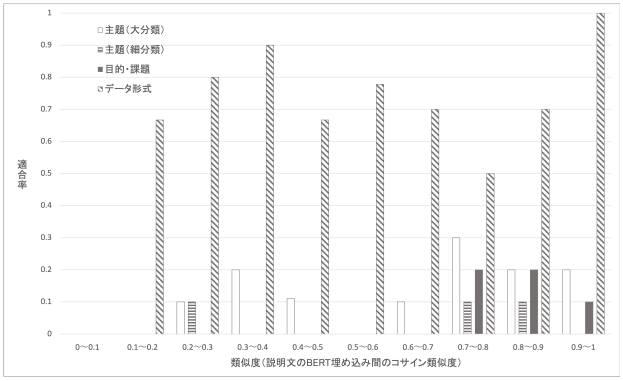

図 4　各評価項目におけるデータセットペア間の適合率（BERT）

5　結論

本研究では，類似データセット発見手法の有効性評価において，データプラットフォームで用いられている詳細なデータセット分類に基づいた 4 種類の評価項目を適用した．結果として，変数ラベル間の Dice 係数に基づく類似度はその数値に比例してデータセットペア間で類似する項目の種類が増加するという傾向が見られた．特に，類似度が非常に高い（0.8〜）領域では，メイントピックなど具体的な項目についての類似が増加することから，この類似度はデータセット間の類似性を適切に反映していると考えられる．そのため，この手法は関連課題や具体的なメイントピックが完全に一致するデータセットを探索するといった状況に特に適していると考えられる．

今後の方針としては，他のデータプラットフォームでの分析や，異なる類似データセット発見手法の評価への適用が考えられる．

参考文献

[1]Magdalena Balazinska, Bill Howe, and Dan Suciu. Data Markets in the Cloud: An Opportunity for the Database Community. PVLDB, Vol. 4, pp. 1482–1485, August 2011.

[2]Florian Stahl, Fabian Schomm, and Gottfried Vossen. Data Marketplaces: An Emerging Species. Databases and Information Systems VIII, pp. 145–158, 2014.

[3]Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. Journal of the Association for Information Science and Technology, Vol. 70, No. 5, pp. 419–432, 2019.

[4]Kathleen Gregory, Andrea Scharnhorst, and Sally Wyatt. Lost or Found? Discovering Data Needed for Research. Harvard Data Science Review, Vol. 2, No. 2, April 2020.

[5]Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 1277–1289, Denver Colorado USA, May 2017. ACM.

[6]Martin Nečaský, Petr Škoda, David Bernhauer, Jakub Klímek, and Tomáš Skopal. Modular framework for similarity-based dataset discovery using external knowledge. Data Technologies and Applications, Vol. 56, No. 4, pp. 506–535, January 2022.

[7]Ulrich Atz. The tau of data: A new metric to assess the timeliness of data in catalogues. In CeDEM14 Conference for E-Democracy and Open Government, Vol. 22, pp. 147–162, 2014.

[8]Auriol Degbelo. Open Data User Needs: A Preliminary Synthesis. In Companion Proceedings of the Web Conference 2020, WWW '20, pp. 834–839, New York, NY, USA, April 2020. Association for Computing Machinery.

[9]Srinivasan Parthasarathy and Mitsunori Ogihara. Clustering Distributed Homogeneous Datasets. In Principles of Data Mining and Knowledge Discovery, Vol. 1910, pp. 566–574, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[10]Srinivasan Parthasarathy and Mitsunori Ogihara. Exploiting Dataset Similarity for Distributed Mining. In José Rolim, editor, Parallel and Distributed Processing, Lecture Notes in Computer Science, pp. 399–406, Berlin, Heidelberg, 2000. Springer.

[11]Ingo Siegert, Ronald Böck, and Andreas Wendemuth. Using a PCA-based dataset similarity measure to improve cross-corpus emotion recognition. Computer Speech & Language, Vol. 51, pp. 1–23, September 2018.

[12]Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A Deeper Look at Dataset Bias. In Domain Adaptation in Computer Vision Applications, Advances in Computer Vision and Pattern Recognition, pp. 37–55. Springer International Publishing, Cham, 2017.

[13]Hiroki Sakaji, Teruaki Hayashi, Kiyoshi Izumi, and Yukio Ohsawa. Verification of Data Similarity using Metadata on a Data Exchange Platform. In 2020 IEEE International Conference on Big Data (Big Data), pp. 4467–4474, December 2020.

[14]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, 2019.

[15]Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word Rotator's Distance. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2944–2960. Association for Computational Linguistics, 2020.