分散文章データ統合解析のための データコラボレーション文章解析

関口拓海 [1]　今倉暁 [2]　櫻井鉄也 [2]
[1] 筑波大学大学院　[2] 筑波大学システム情報系
sekiguchi.takumi.ss@alumni.tsukuba.ac.jp　{imakura,sakurai}@cs.tsukuba.ac.jp

概要

複数機関が保有する個人情報を含むデータの解析のため，プライバシー保護を考慮した学習手法が注目を集めている．近年提案されたデータコラボレーション解析は生データではなく各パーティが独自に次元削減を行った中間表現を共有する手法であるが，文章のような順序情報を持つデータの解析が困難である．本研究では分散文章データの統合解析を目的として，文章を単語の分散表現辞書と順序情報の 2 要素に分解して解析を行うデータコラボレーション文章解析を提案する．文章分類問題の実験の結果，単独機関での解析と比較して約 20 ポイントの分類精度向上（生データを直接共有する場合と比較して 1 ~ 3 ポイント以下の精度低下）を確認した．

1　はじめに

一般に，機械学習では学習用のサンプルデータが多いほど高い精度で学習が可能である．類似したデータを持つ複数の機関の間では，データを共有して多くのサンプルデータを用いた解析を行うことで高精度な解析結果を得ることが期待できる．

しかし，サンプルデータの中にはプライバシー保護のため共有の難しいデータが存在する．医療データなどのセンシティブなデータは直接共有して解析することが困難な場合がある．

そのようなデータを直接共有せず解析するための手法としてデータコラボレーション解析 [1–3] が提案された．データコラボレーション解析は複数のデータ提供パーティが各自の生データに独自に次元削減を行った中間表現を共有する手法で，各パーティ間で生データを秘匿しつつ解析者の下で統合して解析を行える技術である．この技術によりクラウドサービスや外部の解析機関によるデータ解析においてプライバシーを保護しつつ多くのサンプルデータを活用した解析を可能とする．

一方，データコラボレーション解析では文章のようなシーケンスデータを解析する場合，次元削減時に順序情報が失われてしまう欠点がある．文章データでは単語の並び順が文脈を構成しておりデータ解析に影響を与える．解析時に順序情報を活用できないことは学習精度の低下に繋がると考えられる．

そこで，本研究では順序情報を保持して解析を行うために，文章データを単語の分散表現辞書と単語の順序情報に分解して解析を行う手法を提案する．単語の分散表現辞書について解析者に秘匿しつつ解析を行い順序情報を後から復元することで，順序情報を活用した高精度な文章解析を実現する．

また文章分類の実験により性能評価を行い，生データを直接共有する場合に近い十分な分類精度で文章データの解析を行えることを確認する．

2　データコラボレーション解析

データコラボレーション解析は複数パーティに分散したデータを直接共有せずに解析者の下で統合して解析を行うための手法である．複数パーティからなるデータ提供者がそれぞれデータを保有する場合において，各パーティが独自に次元削減を行った中間表現をアンカーデータを介して統合して解析を行う．

データ提供者が計 c パーティ存在し，初期状態としてパーティ i $(1 \le i \le c)$ は n_i 個の m 次元からなるサンプルデータ $X_i \in \mathbb{R}^{n_i \times m}$ を持つと仮定する．

最初に，共有可能な公開データや人工データを用いて作成した r サンプルのデータ $X^{\mathrm{anc}} \in \mathbb{R}^{r \times m}$ をアンカーデータと定義し，各提供者パーティで同一のアンカーデータを共有する．ここでアンカーデータは解析者には公開しない．

次に，X_i, X^{anc} に各パーティ i が個別のマッピング関数 f_i を適用して中間表現 $\tilde{X}_i = f_i(X_i) \in$

$\mathbb{R}^{n_i \times \tilde{m}}$, $\tilde{X}_i^{\text{anc}} = f_i(X^{\text{anc}}) \in \mathbb{R}^{r \times \tilde{m}}$ を作成し，解析者の下に中間表現を集める．マッピング関数 f_i は各パーティが任意に決める関数であり，関数 f_i を持たない解析者は中間表現から元のデータを復元することはできない．マッピング関数は線形または非線形の行単位の変換関数であり，例えば主成分分析による次元削減等が該当する．また，秘匿性向上のため乱数行列を使用した変換を併用しても良い．

その後，中間表現を受け取った解析者はアンカーデータの中間表現 $\tilde{X}_i^{\text{anc}}$ を元にマッピング関数 g_i を作成する．g_i は個別に作成された中間表現を比較可能な形に揃えるための関数で，アンカーデータは元々同じ値であることから $\tilde{X}_i^{\text{anc}}$ に摂動を加えて $g_i(\tilde{X}_i^{\text{anc}}) = g_i(\tilde{X}_i^{\text{anc}})$ $(i \neq j)$ とするための摂動を最小化する次式の全体最小二乗問題 [4] の解として得られる．

$$\min_{E_i, G_i' (i=1,2,\cdots,c), Z(\|Z\|_F=1)} \sum_{i=1}^{c} \|E_i\|_F^2$$
$$s.t.\ (\tilde{X}_i^{\text{anc}} + E_i) G_i' = Z.$$

この式は [4] の手法により次式の特異値分解によって g_i を求めることができる．ただし，$(\tilde{X}_i^{\text{anc}})^\dagger$ は $\tilde{X}_i^{\text{anc}}$ の擬似逆行列，C は正則な正方行列とする．

$$\begin{aligned}
[\tilde{X}_1^{\text{anc}}, \tilde{X}_2^{\text{anc}}, \cdots, \tilde{X}_c^{\text{anc}}] &= [U_1, U_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \\
&\approx U_1 \Sigma_1 V_1^T, \\
G_i &= \arg\min_{G \in \mathbb{R}^{m \times m}} \|Z - \tilde{X}_i^{\text{anc}} G\|_F^2 \\
&= (\tilde{X}_i^{\text{anc}})^\dagger U_1 C.
\end{aligned}$$

最後に，解析者はコラボレーション表現 $\hat{X}_i = \tilde{X}_i G_i \in \mathbb{R}^{n_i \times \hat{m}}$ を作成し，統合したコラボレーション表現を入力として機械学習モデルの学習を行う．

3 提案手法

既存のデータコラボレーション解析ではシーケンスデータの順序情報を保持して学習を行うことができない問題がある．中間表現を作成する際にはマッピング関数で m 次元ベクトルのデータを $\tilde{m}$ ベクトルに変換を行う．このため文章データで同様の変換を行う場合，(単語分散表現) × (単語数) の行列をベクトルに変換する必要があり，その際に単語の順序情報が失われてしまう．

本研究では文章サンプルを単語の分散表現辞書と順序情報の 2 要素に分解してから組み合わせて学習を行う手法を提案する．単語の順序情報は単体では元の文章を復元できないため，秘匿せず直接解析者に渡すことができる．そのため，データコラボレーション解析で分散表現辞書のコラボレーション表現を作成した後に単語の順序情報と組み合わせることで順序情報を保持した統合解析を行う．

複数のパーティからなる提供者が持つ文章データを単一の解析者の下で解析することを考える．解析者は埋め込み層のない RNN をベースとした学習モデルを保有し，単語の順序に従い単語分散表現を入力とすることで解析を行うことができる．

3.1 分散表現辞書の作成

最初に，提供者の間で分散表現辞書を作成し共有する．Word2Vec [5] 等の手法で作成された単語分散表現の事前学習済みモデルを元にして，ランダムに並び替えた順番をインデックスとすることでインデックス-単語-分散表現の間で一意な対応を作成する．作成した対応を元にインデックス-単語の対応 $\boldsymbol{v}$，インデックス-分散表現の対応 $D \in \mathbb{R}^{m \times p}$ を作成しデータ提供者間で共有する．$\boldsymbol{v}, D$ はインデックスに従い行方向に順番にそれぞれ単語, 分散表現を並べることで作成される．なお，m は単語分散表現の実ベクトル次元数を表し，p は事前学習モデルに含まれる単語の総数を表す．[1)]

$$\boldsymbol{v} = (\text{“a“}, \text{“is“}, \text{“pen“}, \text{“this“}),$$
$$D = \begin{bmatrix} 1.2 & 0.5 & 1.8 & 0.2 \\ 1.3 & 0.6 & 1.9 & 0.3 \\ 1.4 & 0.7 & 2.0 & 0.4 \end{bmatrix}.$$

3.2 文章の分解

提供者は共有された対応 $\boldsymbol{v}$ を用いて順序情報を表すシーケンス $\boldsymbol{s}_{ij}$ を作成する．各提供者 i は保有する文章サンプル j を単語ごとに分解して l_{ij} 個の要素を持つ単語列とする．$\boldsymbol{v}$ を用いて単語列の単語をそれぞれ対応するインデックスに置き換えることでシーケンス $\boldsymbol{s}_{ij} \in \{1, 2, \ldots, p\}^{l_{ij}}$ を得る．また，提供者 i の保有するシーケンスをまとめて $\boldsymbol{s}_i$ と表す．

このように作成されたシーケンス $\boldsymbol{s}_{ij}$ は D との演算により学習モデルへの入力を得ることができる．行列 $D = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_p)$ とベクトル $\boldsymbol{s}_{ij} = (s_1, s_2, \ldots, s_{l_{ij}})$ の間において，$D = (\boldsymbol{d}_{s_1}, \boldsymbol{d}_{s_2}, \ldots, \boldsymbol{d}_{s_{l_{ij}}})$ に並び替える演算を $D(:, \boldsymbol{s}_{ij})$ として定義する．このとき $D(:, \boldsymbol{s}_{ij})$ は分散表現による文章を表し，RNN を

1) 文章間の長さを揃えるための空白タグ，事前学習モデルに含まれない単語を表す未知語タグ等も含む．

ベースとした学習モデルへの入力とすることができる．

$$
\begin{aligned}
&(\text{文章})\cdots\cdots\text{“This is a pen“},\\
&\boldsymbol{s}_{ij} = (4, 2, 1, 3),\\
&D(:, \boldsymbol{s}_{ij}) = \begin{bmatrix} 0.2 & 0.5 & 1.2 & 1.8 \\ 0.3 & 0.6 & 1.3 & 1.9 \\ 0.4 & 0.7 & 1.4 & 2.0 \end{bmatrix}.
\end{aligned}
$$

3.3 データコラボレーション文章解析

データコラボレーション解析における $X_i = X^{\text{anc}} = D^\top$ として D のコラボレーション表現を作成する．図 1 に示すように各提供者パーティ i は個別のマッピング関数 f_i を用いて中間表現 $\tilde{D}_i = f_i(D) \in \mathbb{R}^{\tilde{m} \times p}$ を作成する．作成した中間表現を解析者の元へ集め，解析者は中間表現を元にマッピング関数 g_i を作成してコラボレーション表現 $\hat{D}_i \in \mathbb{R}^{\hat{m} \times p}$ を得る．個別に作成された中間表現から元データを復元することはできないため，分散表現辞書 D は解析者に対し秘匿されることになる．[2)]

最後に，コラボレーション表現をシーケンス $\boldsymbol{s}_i$ と組み合わせて解析を行う．提供者はシーケンスを秘匿せずそのまま提供者から解析者に渡す．図 1 に示すように解析者の元で $\hat{D}_i$ と $\boldsymbol{s}_i$ を用いて $\hat{D}_i(:, \boldsymbol{s}_i)$ を作成し解析を行う．$\hat{D}_i(:, \boldsymbol{s}_i)$ は単語の順序に従い行方向に分散表現が並べられている．解析者は RNN をベースとした学習モデルへの入力とすることで単語の順序情報を活用して解析が行える．

3.4 シーケンスからの単語推測

シーケンス $\boldsymbol{s}_i$ は秘匿されないため解析者は単語の使用頻度情報を用いた攻撃が可能である．解析者はシーケンスに出てくる各インデックスの使用回数を数えることで単語の使用頻度の情報を得られる．一般に自然言語では単語ごとに使用頻度に差があるため，データ解析者は使用頻度を比較することで元の単語を推測することができる．

しかし，使用頻度の情報のみでは元の文章の復元は困難である．同程度の使用頻度の単語を正確に分類することは難しく，十分な単語数がある場合の復元は難しい．またこの攻撃への対抗策として，一部の単語を他の単語に置き換えたり空白として扱うといった差分プライバシーの手法によって使用頻度を変化させる手法が考えられる．

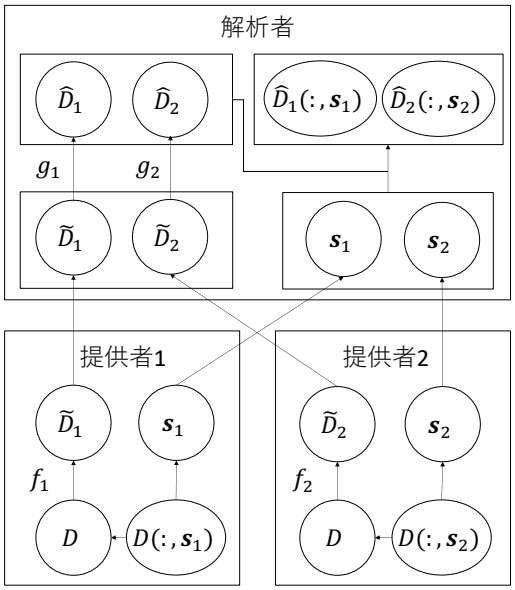

図 1　提案手法による文章解析の概略

4 実験

文章分類の実験によって提案手法の性能評価を行う．提供者が他のパーティとデータを共有せず個別に解析を行う個別解析，生データを秘匿せず直接共有して解析を行う集中解析の 2 つと比較を行う．

4.1 実験設定

実験ではサンプルデータの総数と解析精度の関係を確認する．データ提供者は各パーティ 500 サンプルずつ保有するものとし，パーティ数を 1 から 10 まで変化させることでサンプルデータの総数を 500 から 5000 に変化させる．マッピング関数 f_i では主成分分析による次元削減の後 $[0, 1]$ の一様乱数を要素に持つ乱数行列をかけている．主成分分析では分散表現の事前学習モデル GloVe [6] による 300 次元の分散表現を 128 次元に削減した．解析者は GRU [7] と全結合層からなるニューラルモデルを用いて解析を行う．

サンプルデータには映画レビュー文とニュース記事の 2 種類のデータセットを使用する．1 つ目の実験では IMDB 映画レビューデータセット [8] を使用し，映画のレビュー文を入力として内容が好意的か批判的か判定を行うの 2 値分類を行う．2 つ目の実験ではロイターニュース記事データセット [9] を用いて 46 種類のトピックからニュース記事がどのト

2)　$\hat{D}$ に対し標準化の処理を適用すると学習精度の向上に繋がる場合がある．$\hat{D}$ は g_i を適用する際に D と比較して絶対値が小さい値になりやすいため，ニューラルネットワークによる誤差逆伝播が機能しにくくなることを防ぐ効果がある．

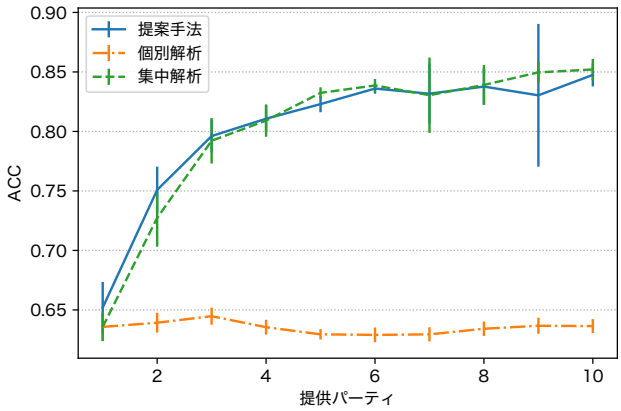

図 2　IMDB 映画レビュー分類

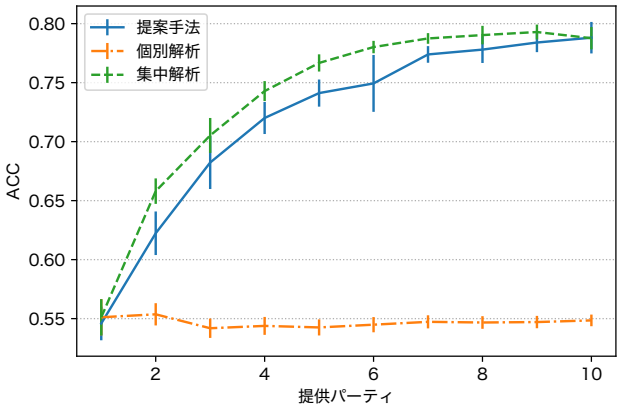

図 3　ロイターニュース記事分類

ピックに対応するかの多値分類を行う．

4.2　結果

図 2 に IMDB 映画レビュー文，図 3 にロイターニュース記事の実験結果を示す．グラフはそれぞれ 10 回試行した平均値を取っており，エラーバーは標準誤差を表している．個別解析では提供パーティを増やしても分類精度（ACC）がほぼ一定の結果となった一方，集中解析では提供パーティを増やし学習に利用するサンプル総数が増えるに従い ACC も向上している．

提案手法の ACC は集中解析に近い値を示し，提供パーティ数に応じて精度が向上していることが読み取れる．提供パーティ数 10 の 5000 サンプル時点では個別解析と比較して 20 ポイント以上高い数値となっており，集中解析との比較では IMDB 映画レビューでは同等の数値，ロイターニュース記事では 1 ~ 3 ポイント程度低い数値となった．

4.3　考察

提案手法により複数パーティのデータを統合して集中解析に近い精度で学習できた．単一パーティのデータしか利用できない個別解析では提供パーティを増やしても ACC は向上しない．一方，全パーティのデータを統合して解析する提案手法と集中解析では提供パーティの増加に応じて ACC も向上しサンプル数に応じて高い精度で解析できている．

また，集中解析と比較して提案手法は映画レビュー分類では同等の精度で学習できているが，ニュース記事分類では 1 ~ 3 ポイント程度低い精度となってしまった．しかし，サンプル数に応じて精度が向上していることから十分なサンプル数を用意すれば実用に足る十分な精度で学習できると考えられる．

提案手法が高い精度で学習できていることは，文章の順序情報を保持して学習できていることが大きな要因であると考察する．提案手法では単語の並び順を考慮して学習ができるため前後の文脈を踏まえて分類を行えるため集中解析に近い精度を出せたと考えられる．

提案手法と集中解析の精度の違いは分散表現辞書のコラボレーション表現作成による違いだと考察する．分散表現辞書は中間表現の作成時に次元削減と乱数列の積をとっており，次元削減により失われる情報やマッピング関数 g_i によって各パーティの乱数列の違いの影響を打ち消せなかった部分が精度に現れたと考えられる．

5　おわりに

本研究では文章を分散表現辞書とシーケンスで表現し，順序情報を活用しながらプライバシー保護を考慮して学習を行う手法を提案した．文章分類の実験を行い，提案手法により文章の順序情報を活用して集中解析に近い精度で学習できることを確認した．

一方，提案手法による解析時は解析者の元に単語の使用頻度の情報を渡す事になる欠点を抱えている．使用頻度から元の文章を復元する攻撃の成功率の検証，差分プライバシーによる対策を行う場合の精度低下の検証は今後の課題としたい．

謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の委託業務（JPNP18010）の結果得られたものです.

参考文献

[1] Akira Imakura, Tetsuya Sakurai. Data Collaboration Analysis Framework Using Centralization of Individual Intermediate Representations for Distributed Data Sets. **ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering**, Vol. 6, No. 2, p. 04020018, 2020.

[2] Akira Imakura, Xiucai Ye, and Tetsuya Sakurai. Collaborative Data Analysis: Non-model Sharing-Type Machine Learning for Distributed Data. In Hiroshi Uehara, Takayasu Yamaguchi, and Quan Bai, editors, **Knowledge Management and Acquisition for Intelligent Systems**, Vol. 12280 of **PKAW 2021. Lecture Notes in Computer Science**, pp. 14–29, Cham, 2021. Springer International Publishing.

[3] Akira Imakura, Ryoya Tsunoda, Rina Kagawa, Kunihiro Yamagata, Tetsuya Sakurai. DC-COX: Data collaboration Cox proportional hazards model for privacy-preserving survival analysis on multiple parties. **Journal of Biomedical Informatics**, Vol. 137, p. 104264, 2023.

[4] Shinji Ito and Kazuo Murota. An Algorithm for the Generalized Eigenvalue Problem for Nonsquare Matrix Pencils by Minimal Perturbation Approach. **SIAM Journal on Matrix Analysis and Applications**, Vol. 37, No. 1, p. 409–419, 2016.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013.

[6] Jeffrey Pennington and Richard Socher and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In **Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1532–1543, 2014.

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014.

[8] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[9] Reuters-21578,Distribution 1.0. http://www.daviddlewis.com/resources/testcollections/reuters21578.