7D-01

テンソル分解の著者名寄せへの応用と潜在変数を持つモデルとの比較

蔵川　圭†　　馬場　康維‡

国立情報学研究所†　統計数理研究所‡

1. はじめに

電子図書館における代表的な問題であり，計算機が使われる 1950 年代から指摘されつつも未だに十分な解決を見ていない著者の名寄せの問題を取り扱う．書誌データに記述される著者名だけでは著者を特定することは不十分であるため，アルゴリズムで書誌データを著者ごとに分類する際には著者を潜在変数に割り当てたモデルを構築し，クラスタリングすることがよく行われる[1]．本報では，近年様々な情報分析や予測において取り上げられているテンソル分解（たとえば，[2][3]）を応用検討することを主眼とし，実験では潜在変数を持つモデルの一つである LDA(Latent Dirichlet Allocation)との比較を行う．

2. テンソルとテンソル分解

テンソルの数学的定義は次のようである[4]．p個の任意のベクトル$\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$に対して実数値

$$\mathbf{T}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$$

を対応させる$\mathbf{T}$が存在して，それぞれのベクトル変数について線型性

$$\begin{cases} \mathbf{T}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r + \boldsymbol{v}'_r, \ldots, \boldsymbol{v}_p) \\ \quad = \mathbf{T}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \ldots, \boldsymbol{v}_p) + \mathbf{T}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}'_r, \ldots, \boldsymbol{v}_p) \\ \mathbf{T}(\boldsymbol{v}_1, \ldots, k\boldsymbol{v}_r, \ldots, v_p) = k\mathbf{T}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r, \ldots, \boldsymbol{v}_p) \end{cases}$$

が成り立つ時，関数$\mathbf{T}$をp階のテンソルといい，pをそのテンソルの階数という．

テンソル$\mathbf{T}$は，いま，直交基底$\Sigma\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$をとるとき，$\mathbf{T}(\boldsymbol{e}_i, \boldsymbol{e}_j) = T_{ij}$ $(i, j = 1,2,3)$とおくと，

$$\mathbf{T}(\boldsymbol{u}, \boldsymbol{v}) = \mathbf{T}\left(\sum_{i=1}^{3} u_i \boldsymbol{e}_i, \sum_{j=1}^{3} v_j \boldsymbol{e}_j\right) = \sum_{i=1}^{3}\sum_{j=1}^{3} \mathbf{T}(u_i \boldsymbol{e}_i, v_j \boldsymbol{e}_j)$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} v_j u_i \mathbf{T}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sum_{i=1}^{3}\sum_{j=1}^{3} T_{ij} u_i v_j$$

となるので，T_{ij}を知れば，$\mathbf{T}(\boldsymbol{u}, \boldsymbol{v})$の性質がわかることになる．$(T_{ij})$を 2 階のテンソル$\mathbf{T}$の基底$\Sigma$に関する成分という．$p$階のテンソルの成分も同様に定義できる．テンソルの成分は計算機ではp次元配列上の要素に対応させることが可能であり，p次元配列がテンソル空間を規定する．

テンソル空間を用いた応用は，物理学における質点系，流体，弾性体の力学や電磁気学の中でベクトル解析の延長として現れるが，統計学におけるテンソルは，p次元空間を対象とした多変量解析の一手法として登場する．解析のアプローチの一つであるテンソル空間の分解は，2 次元行列を対象とした SVD や PCA と同様に，多次元空間の次元圧縮や潜在特徴量の把握に利用されている．分解の方法は，すでにいくつも発見されており，CP(CANDECOMP/PARAFAC)分解や Tucker 分解がとりあげられることが多い．テンソル分解の解説は，[5]に詳しい．

3. 著者クラスタリングの方法

3.1. 著者推定問題の定式化

テンソル分解によって多次元空間の潜在特徴量の解析が可能であるから，これを書誌レコードから著者を推定する問題に適用することを考える．

前提とする書誌レコードは図 1 に例示するように一般の書誌項目で構成されているとする．

```
CID,AFID,JNAME,ENAME,YNAME,JAFF,EAFF,YEAR,CO-AUTH,TITLE,JRNL
26,B-10002920439-3-CJP,松本健一,Matsumoto Ken-ichi,マツモト ケンイチ
,北大・院薬・分子生物,Dep. Mol. Bio. Grad. Sch. Of Pharm. Sci. Hokkaido Univ.,
1998,"生田 智樹,有賀 寛芳,松本 健一",
細胞外マトリックス・テネイシンxと相互作用する分子の探索,
日本分子生物学会年会プログラム・講演要旨集
```

図 1 書誌レコードの例

書誌に記載される共著者の一人を指定したレコードをここでは著者フラグメント(AF)と呼ぶ．著者を推定する問題は，この著者フラグメントを同一著者のクラスターとして構成する分類問題に置き換えることができる(図 2)．

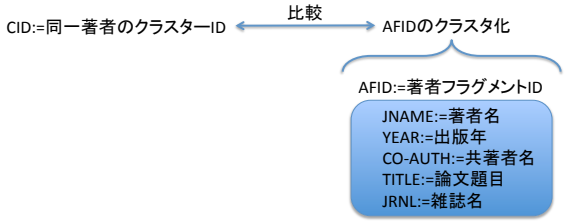

図 2 著者フラグメントのクラスタ化

3.2. CP 分解を応用した著者特徴ベクトルの導出と k-means によるクラスタリング

分類問題の解法は，k-means，Agglomerative clustering，DBSCAN，Affinity propagation，Spectral clustering などが知られている．これらは比較対象とする 2 点間の距離と解法独自のクラスター数などのパラメータをあらかじめ与えることで，点をアルゴリズムによって機械的に分類する．類似度は距離の逆数であるから，

Applying tensor decomposition for the name disambiguation problem and comparative study of models with latent variables
† Kei Kurakawa, National Institute of Informatics
‡ Yasumasa Baba, The Institute of Statistical Mathematics

距離を類似度と読み替えても良い．

著者フラグメントを同一著者で分類するためには，著者フラグメントを点に対応させ，著者同一性を表す類似度を定義できれば良い．類似度の定義として，ここでは次の操作を行う．まず，著者同一性判定に寄与すると思われる書誌属性の類似度を 3 階のテンソルへマッピングする（図 3，表 1）．ここでのテンソルを$\mathcal{X}$とおく．

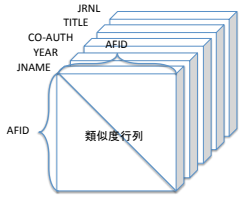

図 3 テンソルスライス上の書誌属性類似度行列

表 1 書誌属性ごとの著者フラグメント類似度の定義

属性	説明
JNAME	著者名の一致 1：一致のとき，　0：不一致のとき
YEAR	出版年の一致 1：一致のとき，　0：不一致のとき
CO-AUTH	共著者名(JNAME を除く)の一致数
TITLE	形態素(名詞，未知語)の一致数
JRNL	形態素(名詞，未知語)の一致数

次に，CP 分解を行い，rank-one テンソルからなるR個の因子に分解する．

$$\mathcal{X} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r$$

$\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r$はテンソルのモード 1，2，3 に対応し，それぞれ AFID，AFID，属性名を要素名としてもつ．因子化された$\boldsymbol{a}_r$ $(r = 1, \ldots, R)$で構成される行列$\mathbf{A} = [\boldsymbol{a}_1 \quad \cdots \quad \boldsymbol{a}_R]$の行ベクトルは著者フラグメントの潜在特徴を表している．ここでは，因子数Rを行列$\mathbf{A}$の列の値の分散が大きくなるように設定し，k-means を用いて行列$\mathbf{A}$を入力として AFID を分類する．

4. 実験

4.1. 方法

実験を遂行するにあたって，公開されている既存のツールを用いた．テンソルの演算のために scikit-tensor，k-means は scikit-learn を用いた．分類手法の比較のために，LDA を用いた．著者クラスターと著者フラグメントの関係を LDA のトピックとドキュメントの関係に割り当て，ドキュメントに対して最尤のトピックを分類結果とする分類を行った．実装は Mallet を用いた．さらに，ランダムに分類することも行った．以上の分類手法の性能を比較するために，Purity, Inverse-purity 指標[6]を用いた．

4.2. データセット

実験に用いたデータセットは 2 種である．それぞれ，同一の姓名をもつ著者フラグメントの集合であり，正解ラベル（クラスターID）が付与されている．サイズは，それぞれ 1568，1121 であり，正解クラスター数は，48，119 である．クラスターサイズを図 4 に示す．

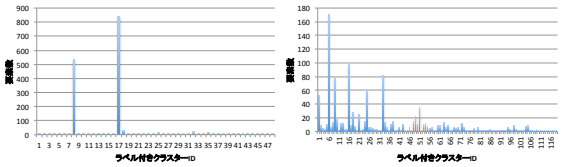

図 4 Matu-ci データセット（左）と Koba-ci データセット（右）の著者クラスターサイズ

4.3. 結果

実験結果を表 2 に示す．ランダムにクラスターID を付与する方法よりも，圧倒的にテンソルを用いた方法と LDA を用いた方法は良い結果を残した．しかしながら，LDA の方が最も良い結果となった．

表 2 実験結果の比較

	Tensor CP-decomp. and K-means		LDA		Random labeling (baseline)	
Dataset	Purity	Inv.-	Purity	Inv.-	Purity	Inv.-
Matu-ci	0.8616	0.1888	**0.8948**	**0.2251**	0.5434	0.0721
Koba-ci	0.5834	0.3943	**0.7565**	**0.4835**	0.2194	0.1320

5. 考察と展望

潜在変数を持つ分類のモデルは，著者を潜在変数に結びつけることで著者同一性判定に有効であることがわかった．テンソルを用いた方法は多次元データを空間上に素直に表現することが可能であり，演算の工夫の仕方で多様な相関関係を導き出すことを可能とする．今後は，テンソル空間上での演算を新たに考案する．

謝辞
実験で用いたデータセットを提供いただきました国立情報学研究所の相澤彰子氏に感謝申し上げます．

参考文献

[1] Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. ACM SIGMOD Record, 41(2), 15. doi:10.1145/2350036.2350040

[2] Nakatsuji, M., Fujiwara, Y., Toda, H., Sawada, H., Zheng, J., & Hendler, J. A. (2014). Semantic data representation for improving tensor factorization. In Proceedings of the National Conference on Artificial Intelligence (Vol. 3, pp. 2004-2012).

[3] Matsubara, Y., Sakurai, Y., Panhuis, W. G. van, & Faloutsos, C. (2014). FUNNEL: automatic mining of spatially coevolving epidemics. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:10.1145/2623330.2623624

[4] 田代嘉宏，テンソル解析（基礎数学選書 23），1981

[5] Kolda, T. G., & Bader, B. W. (2009). Tensor Decompositions and Applications. SIAM Review, 51(3), 455-500. doi:10.1137/07070111X

[6] Artiles, J., Gonzalo, J., & Sekine, S. (2007). The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007) (pp. 64-69).