5M-05

データコラボレーション解析における統合表現の最適化と加重法

川上雄大† 高野祐一‡ 今倉暁‡

†筑波大学大学院 システム情報工学研究群 ‡筑波大学 システム情報系

1 はじめに

近年，IT 技術の発達により，多くの立場からデータを解析することが可能になった．しかし，解析の精度を上げるためには十分な量のデータが必要になる．機関内に十分な量のデータがない状況では，各機関 i に分散した元データ $\boldsymbol{X}_i$ を 1 つの機関内で解析する個別解析よりも，全機関の元データを 1 箇所に集めて解析する集中解析の方が解析精度は高い．

しかし，元データを集約するには機密情報保護が問題となる．そこで，データコラボレーション解析（DC 解析）が Imakura and Sakurai [2] によって提案された．DC 解析では，各機関において中間関数 f_i で元データを中間表現 $\tilde{\boldsymbol{X}}_i$ へと抽象化し，機密情報が保護されている中間表現を全機関から 1 箇所に集約する．中間表現について，元データの機密情報が保護されることは Imakura et al. [1] によって証明されている．中間表現を集約・解析することで，元データを直接集約しなくても，集中解析と同様に個別解析よりも解析精度が向上することは注目すべき点である．

ただし，中間表現はそのまま他の機関の中間表現と一緒に計算をすることはできない．そのため，中間表現を一緒に計算できる状態 (統合表現 $\hat{\boldsymbol{X}}_i$) に変換する統合関数 g_i が必要になる．既存手法 [2] では，統合関数最適化問題を最小摂動問題として定式化し，統合関数を求めていた．しかし，従来の統合関数最適化問題では，零行列の解を排除するために強い直交性の制約を加えていることなどが課題となっていた．

本研究では，それらの課題を解決するために，列ベクトルごとに分割した最適化問題の定式化と 2 種類の解法を提案する．また数値実験では，提案手法と既存手法の性能を比較する．

Optimization and weighting methods of collaboration representations in the data collaboration analysis

Yuta KAWAKAMI†, Yuichi TAKANO†and Akira IMAKURA‡

†Degree Programs in Systems and Information Engineering, University of Tsukuba

‡Institute of Systems and Information Engineering, University of Tsukuba

2 統合関数最適化問題

先行研究 [2] で提案されている統合関数最適化問題の定式化と解法について紹介する．

統合関数を得るために，アンカーデータと呼ばれる共通データ $\boldsymbol{X}^{\mathrm{anc}}$ が全機関に共有される．アンカーデータはランダムなダミーデータで構築される．また，アンカーデータは元データと同じ各機関の中間関数 f_i で中間表現 $\tilde{\boldsymbol{X}}_i^{\mathrm{anc}}$ に抽象化され，1 箇所に集約される．

アンカーデータが全機関に共有された同じ値を持つデータであることを用いて，アンカーデータの統合表現 $\hat{\boldsymbol{X}}_i^{\mathrm{anc}}$ が全機関で近似の値を取るように統合関数は最適化される．ここで，統合関数 g_i を行列 $\boldsymbol{G}_i$ による線形変換，統合表現の次元数を $\hat{m}$，DC 解析でデータ共有を行う機関数を N と置く．このとき，アンカーデータの統合表現の目標行列 $\boldsymbol{Z} \in \mathbb{R}^{r\times\hat{m}}$ に対して，全てのアンカーデータの統合表現 $\hat{\boldsymbol{X}}_i^{\mathrm{anc}} = \tilde{\boldsymbol{X}}_i^{\mathrm{anc}}\boldsymbol{G}_i$ が近づくように次の定式化を行う．

$$\min_{\boldsymbol{G}_i(i\in[N]),\boldsymbol{Z}} \sum_{i=1}^{N} \|\boldsymbol{Z} - \tilde{\boldsymbol{X}}_i^{\mathrm{anc}}\boldsymbol{G}_i\|_{\mathrm{F}}^2 \tag{1}$$

ただし，この問題はそのまま解くと，Z と G_i が零行列になってしまう．そこで，最小摂動問題を用いて，単位行列 $\boldsymbol{I}$ に対して，$\boldsymbol{Z}^\top\boldsymbol{Z} = \boldsymbol{I}$ の下で限定的に問題を解く．まず，アンカーデータの中間表現 $\tilde{X}_i^{\mathrm{anc}}$ を組み合わせた行列の特異値分解，

$$\left(\tilde{\boldsymbol{X}}_1^{\mathrm{anc}}\ \tilde{\boldsymbol{X}}_2^{\mathrm{anc}}\ \ldots\ \tilde{\boldsymbol{X}}_N^{\mathrm{anc}}\right) = (\boldsymbol{U}_1\,\boldsymbol{U}_2)\,\Sigma\boldsymbol{V}^\top \tag{2}$$

によって，行列 $\boldsymbol{U}_1 \in \mathbb{R}^{r\times\hat{m}}$ を求める．

次に目標行列を $\boldsymbol{Z} = \boldsymbol{U}_1$ と置く．新たに定めた目標行列を用いて次の線形最小二乗問題を解くことによって，行列 $\boldsymbol{G}_i$ は機関 i ごとに求まる．

$$\boldsymbol{G}_i = \arg\min_{\boldsymbol{G}} \|\boldsymbol{Z} - \tilde{\boldsymbol{X}}_i^{\mathrm{anc}}\boldsymbol{G}\|_{\mathrm{F}}^2 = (\tilde{\boldsymbol{X}}_i^{\mathrm{anc}})^\dagger\boldsymbol{U}_1 \tag{3}$$

ただし，$(\tilde{\boldsymbol{X}}_i^{\mathrm{anc}})^\dagger$ は $\tilde{\boldsymbol{X}}_i^{\mathrm{anc}}$ の一般化逆行列を示す．

以上のように最小摂動問題に問題を置き直して解くことで，中間表現を統合する行列 $\boldsymbol{G}_i$ を推定できる．

3 提案手法

3.1 交互最適化による統合関数最適化

1つ目の提案手法では，まず目標行列 $\boldsymbol{Z}$ と行列 $\boldsymbol{G}_i$ を次のように列ベクトルごとに分解する．

$$\boldsymbol{Z} = (\boldsymbol{z}_1 \dots \boldsymbol{z}_j \dots \boldsymbol{z}_{\hat{m}}) \tag{4}$$

$$\boldsymbol{G}_i = (\boldsymbol{g}_{i1} \dots \boldsymbol{g}_{ij} \dots \boldsymbol{g}_{i\hat{m}}) \tag{5}$$

このように行列を列ベクトルごとに分解することで，$j \in [\hat{m}]$ ごとに，式 (1) の行列の残差の最小化問題を次の列ベクトルの残差の最小化問題に置き換えることができる．

$$\begin{aligned} &\min_{\boldsymbol{g}_{ij}, \boldsymbol{z}_j} F(\boldsymbol{z}_j, \boldsymbol{g}_{ij}) = \sum_{i=1}^{N} \|\boldsymbol{z}_j - \tilde{\boldsymbol{X}}_i^{\mathrm{anc}} \boldsymbol{g}_{ij}\|_2^2 \\ &\text{s.t. } C(\boldsymbol{z}_j) = \boldsymbol{z}_j^\top \boldsymbol{z}_j - 1 = 0 \end{aligned} \tag{6}$$

問題を j 列ごとに分解し，$\boldsymbol{z}_j$ ごとにノルム制約を加えることで，交互最適化の各ステップで，それぞれ解析解が求まる．交互最適化の各ステップは，

1. $\boldsymbol{z}_j$ を固定した際の $\boldsymbol{g}_{ij}$ の最適解推定
2. $\boldsymbol{g}_{ij}$ を固定した際の $\boldsymbol{z}_j$ の最適解推定

の2ステップである．1. は最小二乗法，2. はラグランジュの未定乗数法により最適解推定を行う．$\boldsymbol{z}_j$ を初期化し，上記の2ステップを繰り返して目的関数の値が収束した段階で統合関数が得られる．

3.2 一般化固有値問題による統合関数最適化

これまでの最適化問題では，統合表現と目標行列を近づけるように定式化を行なっていた．ここでは，統合表現同士を近づけるように問題を置き換える．

任意の2つの機関を i，i' と置く．このとき，$j \in [\hat{m}]$ ごとに，全統合表現の差を目的関数，統合表現のノルム制約を制約として次の定式化を行う．

$$\begin{aligned} \min_{\boldsymbol{g}_{ij}} F(\boldsymbol{v}_j) &= \sum_{i=1}^{N} \sum_{i'=1}^{N} \|\tilde{\boldsymbol{X}}_i^{\mathrm{anc}} \boldsymbol{g}_{ij} - \tilde{\boldsymbol{X}}_{i'}^{\mathrm{anc}} \boldsymbol{g}_{i'j}\|_2^2 \\ &= \boldsymbol{v}_j^\top \boldsymbol{A}_{\tilde{X}} \boldsymbol{v}_j \\ \text{s.t. } C(\boldsymbol{v}_j) &= \sum_{i=1}^{N} \|\tilde{\boldsymbol{X}}_i^{\mathrm{anc}} \boldsymbol{g}_{ij}\|_2^2 - 1 \\ &= \boldsymbol{v}_j^\top \boldsymbol{B}_{\tilde{X}} \boldsymbol{v}_j - 1 = 0 \end{aligned} \tag{7}$$

ただし，$\boldsymbol{A}_{\tilde{X}}$，$\boldsymbol{B}_{\tilde{X}}$ はある行列，$\boldsymbol{v}_j$ は全機関の $\boldsymbol{g}_{ij}$ を全て繋げたベクトルを表す．

ここでラグランジュの未定乗数法を用いると，次の一般化固有値問題に帰着する．

$$\boldsymbol{A}_{\tilde{X}} \boldsymbol{v}_j = \lambda_j \boldsymbol{B}_{\tilde{X}} \boldsymbol{v}_j \quad (\boldsymbol{v}_j^\top \boldsymbol{B}_{\tilde{X}} \boldsymbol{v}_j = 1) \tag{8}$$

ここで，目的関数 $F(\boldsymbol{v}_j)$ の極値と一般化固有値 λ_j は等しい．$\boldsymbol{v}_j$ は全機関の $\boldsymbol{g}_{ij}$ を全て繋げたベクトルであるので，小さい順に並べた λ_j に対応する $\boldsymbol{v}_j$ より統合関数が求まる．また $j \in [\hat{m}]$ であるので，λ_j は小さい順に $\hat{m}$ 番目 ($\lambda_1 < \lambda_2 < \ldots < \lambda_{\hat{m}}$) まで求める．

加重法

これまで，次元数 $\hat{m}$ を事前に定めておき，列 $j \in [\hat{m}]$ のベクトルを求めて統合関数を構築していた．そこで，一般化固有値問題による統合関数最適化を対象として，加重法による統合関数の効果的構築手法を提案する．式 (7) の統合関数最適化問題の目的関数の値は一般化固有値 λ_j に一致する．このことから，一般化固有値 λ_j は統合関数の精度の悪さを示す．これが加重法の前提となる考えである．

統合表現 $\hat{X}_i$ の列ベクトル $\hat{\boldsymbol{x}}_{ij}$ は，統合表現の j 番目の特徴量を示しており，$\hat{\boldsymbol{x}}_{ij}$ は統合関数 $\boldsymbol{g}_{ij}$ によって生成される．このとき，統合関数 $\boldsymbol{g}_{ij}$ の精度が悪いと，対応する特徴量 $\hat{\boldsymbol{x}}_{ij}$ によって解析時に悪影響が及ぶ．そこで加重法では，次に示す加重関数を用いることで，$\hat{\boldsymbol{x}}_{ij} \leftarrow w(\lambda_j)\hat{\boldsymbol{x}}_{ij}$ のように，統合表現の解析前に悪影響の大きい特徴量のスケールを減衰させる．

$$w(\lambda_j) = \exp\left(-\frac{\lambda_j - \lambda_1}{\lambda_{\hat{m}} - \lambda_1}\right) \tag{9}$$

4 数値実験

統合関数最適化の既存手法と提案手法の比較を行い，提案手法の有効性を検証する．実験の詳細は当日述べる．

参考文献

[1] Akira Imakura, Anna Bogdanova, Takaya Yamazoe, Kazumasa Omote and Tetsuya Sakurai. “Accuracy and privacy evaluations of collaborative data analysis.” The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence. PPAI-21, 2021.

[2] 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 6.2 (2020): 04020018.