社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

関係データの直積空間への埋め込みによる可視化

宮崎　一希†　渡辺　龍二†　古川　徹生†

† 九州工業大学大学院生命体工学研究科　〒808–0196 福岡県北九州市若松区ひびきの 2–4
E-mail: †{miyazaki.kazuki658,watanabe.ryuji717}@mail.kyutech.jp, ††furukawa@brain.kyutech.ac.jp

あらまし　本研究の目的は関係データのモデリングおよび可視化を行う手法の開発である．関係データは複数ドメインのオブジェクト組から観測されたデータである．ドメインごとに対応する潜在空間へオブジェクトを埋め込む，すなわち潜在空間の直積空間へデータを埋め込むことが提案手法の目的である．本稿ではカーネル平滑化を用いたノンパラメトリックな多様体モデリングを提案する．
キーワード　関係データ解析, 可視化, カーネル平滑化, 積多様体

Visualization of relational data by Embedding to Direct product space

Kazuki MIYAZAKI†, Ryuji WATANABE†, and Tetsuo FURUKAWA†

† Graduate School of Life Science and System Engineering, Kyushu Institute of Technology　Hibikino 2–4, Wakamatu–ku, Kitakyushu–shi, Fukuoka, 808–0196 Japan
E-mail: †{miyazaki.kazuki658,watanabe.ryuji717}@mail.kyutech.jp, ††furukawa@brain.kyutech.ac.jp

Abstract　The aim of this work is to develop a modeling method of relational data. Relational data is a dataset observed obtained from object pairs belonging to several domains. The task of the proposed method is to embed the objects into the direct product latent spaces, each of which is corresponding to the domain. In this work, we employed the kernel soother to represent the product manifold, and it estimates the latent variables by a non-parametric unsupervised manner.
Key words　Relational data analysis, visualization, kernel smoother, product manifold

1. はじめに

本研究の目的は関係データの解析およびモデリング手法の開発である．関係データとは複数のオブジェクトの関係を表現したデータのことで, 複数ドメインのオブジェクトの組み合わせに対して 1 つのデータが観測される [1]．関係データの典型例は EC サイトの商品評価データであり，これは「顧客（ユーザー）」が「商品（アイテム）」に対して評価を与えたものである．すなわちユーザーと商品の組に対してひとつの評価データが与えられる．このような商品評価データを解析できれば，ユーザーに関するクラスタやその嗜好傾向，アイテムに関するクラスタや顧客層などの知識発見が可能になる．またモデル化できれば新規顧客の嗜好予測や，新規商品のターゲット層予測などに用いることもできる．

関係データから知識発見するには二種の解析が必要になる．第一はドメインごとの解析であり，商品評価データの場合はユーザー同士もしくは商品同士の関係を解析することに該当する．第二はドメイン間の関係の解析であり，商品評価データの場合はどのようなユーザーがどのような商品を高く評価するかの解析に該当する．本研究では特にオブジェクトを低次元の潜在空間へ写像することで関係データを可視化することを目的とする．

一方，関係データのモデリングで問題となるのは，観測データの網羅性である．もしすべてのオブジェクト組についてデータが観測されているならば，関係データは行列もしくはテンソルとして表現される．このような場合は行列分解やテンソル分解などの手法が有用である [2]．しかしながら，データが観測されるのは一部のオブジェクト組のみであることがしばしばである．このような場合，行列・テンソル分解によるアプローチでは欠損値補完などを行う必要がある．また非線形モデリングの場合はモデリングの表現方法も問題になる．なぜなら潜在空間の直積空間をドメインとする非線形写像の表現をするには，基底関数の数が直積空間の次元のべき乗で増加するからである．また関係データは中規模サイズのデータであっても計算コストが増大しがちで，効率の良い計算方法が必要となる．

本研究では以下の機能を持つ手法の提案をめざしている．

- 関係データをドメインごとの潜在空間の直積空間へ埋め込むことで，関係データの解析とモデリングを可能にする．
- 潜在空間の次元をドメインあたり 2 次元にすることで，ドメインごとに可視化し，ドメイン内解析を可能にする．

- 潜在空間から観測空間への埋め込み写像をモデリングし，新規データへの予測を可能にする．その際，計算コストが潜在空間の次元に依存しないノンパラメトリックなモデリングを行う．
- 潜在空間から観測空間への写像をモデリングすることで，ドメイン間解析を可能にする．
- 関係データをテンソルとして扱う必要がなく，欠損補完という概念を必要としないモデリング法を実現する．

2. 問題設定

K 個のオブジェクト組から得られた関係データを K 項関係データと呼ぶ．話を簡単にするため本稿では 2 項関係データの場合について述べるが，任意の K 項データへの一般化は容易である．

今，$\Omega^{(1)} = \left\{\omega_i^{(1)}\right\}_{i=1}^{I}$ を第 1 ドメインのオブジェクト集合，$\Omega^{(2)} = \left\{\omega_j^{(2)}\right\}_{j=1}^{J}$ を第 2 ドメインのオブジェクト集合とする．また観測データを $\mathbf{X} = (\mathbf{x}_n)_{n=1}^{N}$，$\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^{D_\mathcal{X}}$ とし，$\mathbf{x}_n$ はオブジェクト組 $\left(\omega_{i(n)}^{(1)}, \omega_{j(n)}^{(2)}\right)$ から観測されたとする．通常の多様体モデリングでは観測空間 $\mathcal{X}$ が高次元であると仮定するが，関係データの場合にはその仮定を必要とせず，$D_\mathcal{X} = 1$ であってもかまわない．

一方，$\mathcal{Z}^{(m)} \subseteq \mathbb{R}^{D_\mathcal{Z}}$ を $\Omega^{(m)}$ に対応する低次元の潜在空間とする．われわれの第一の目的はオブジェクト集合 $\Omega^{(1)}, \Omega^{(2)}$ をそれぞれ $\mathcal{Z}^{(1)}, \mathcal{Z}^{(2)}$ に写像することである．すなわち

$$\varphi^{(1)}\colon \Omega^{(1)} \to \mathcal{Z}^{(1)}\colon \omega_i^{(1)} \mapsto \mathbf{z}_i^{(1)}$$
$$\varphi^{(2)}\colon \Omega^{(2)} \to \mathcal{Z}^{(2)}\colon \omega_j^{(2)} \mapsto \mathbf{z}_j^{(2)}$$

である．特に $D_\mathcal{Z} = 2$ の場合は各ドメインの潜在空間上にオブジェクトが配置されたマップとして可視化することができる．また第二の目的は潜在空間の直積空間から観測空間への写像 $f\colon \mathcal{Z}^{(1)} \times \mathcal{Z}^{(2)} \to \mathcal{X}\colon (\mathbf{z}^{(1)}, \mathbf{z}^{(2)}) \mapsto \mathbf{x}$ によって関係データを $\mathbf{x}_n \simeq f\left(\mathbf{z}_{i(n)}^{(1)}, \mathbf{z}_{j(n)}^{(2)}\right)$ とモデル化することである．ここで f は関数空間 $\mathcal{H}_{\mathrm{RKHS}}$ に属する滑らかな連続写像とする．このとき直積潜在空間 $\mathcal{Z}^{(1)} \times \mathcal{Z}^{(2)}$ は $\mathcal{Z}^{(1)} \times \mathcal{Z}^{(2)} \times \mathcal{X}$ 内に多様体として埋め込まれる．

以上よりわれわれの目的は $\mathbf{x}_n$ と $f(\mathbf{z}_{i(n)}^{(1)}, \mathbf{z}_{j(n)}^{(2)})$ の二乗誤差をなるべく小さくし，かつなるべく滑らかな f および潜在変数 $\mathbf{Z}^{(1)} = \left(\mathbf{z}_i^{(1)}\right), \mathbf{Z}^{(2)} = \left(\mathbf{z}_j^{(2)}\right)$ を推定することである．

3. 目的関数

提案手法では目的関数を次式で定義する．

$$\begin{aligned} &F\left[\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, f\right] \\ &:= -\frac{\beta}{2}\sum_{n=1}^{N}\iint k(\boldsymbol{\zeta}^{(1)}, \mathbf{z}_{i(n)}^{(1)})\, k(\boldsymbol{\zeta}^{(2)}, \mathbf{z}_{j(n)}^{(2)}) \left\|\mathbf{x}_n - f\left(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}\right)\right\|^2 d\boldsymbol{\zeta}^{(1)}\, d\boldsymbol{\zeta}^{(2)} \\ &\qquad - \sum_{i=1}^{I}\log p(\mathbf{z}_i^{(1)}) - \sum_{j=1}^{J}\log p(\mathbf{z}_j^{(2)}) \end{aligned} \quad (1)$$

ここで $k(\mathbf{z}, \mathbf{z}')$ は平滑化カーネルのカーネル関数であり，本研究ではガウス関数を用いる．また $p(\mathbf{z}^{(1)}), p(\mathbf{z}^{(2)})$ は $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$ の事前分布である．なお本研究では $p(\mathbf{z}^{(m)})$ をコンパクトな潜在空間（単位正方形 $[0,1]^{D_\mathcal{Z}}$）上の一様分布とし，以下の記述では事前分布の項を省略する．しかし非コンパクトな潜在空間上の事前分布（たとえばガウス事前分布）を用いることも可能である．

さて目的関数 (1) を f で変分して極小点を求めると次式が得られる．

$$f\left(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)} \mid \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}\right) = \frac{1}{K(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)})}\sum_{n=1}^{N} k(\boldsymbol{\zeta}^{(1)}, \mathbf{z}_{i(n)}^{(1)})\, k(\boldsymbol{\zeta}^{(2)}, \mathbf{z}_{j(n)}^{(2)})\, \mathbf{x}_n \quad (2)$$

ここで

$$K(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}) := \sum_{n=1}^{N} k(\boldsymbol{\zeta}^{(1)}, \mathbf{z}_{i(n)}^{(1)})\, k(\boldsymbol{\zeta}^{(2)}, \mathbf{z}_{j(n)}^{(2)})$$

である．(2) を (1) に代入することで，目的関数から f を消し，潜在変数 $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$ のみの形にすることができる．

$$\begin{aligned} F\left[\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}\right] = -\frac{\beta}{2}\sum_{n=1}^{N}\iint & k(\boldsymbol{\zeta}^{(1)}, \mathbf{z}_{i(n)}^{(1)})\, k(\boldsymbol{\zeta}^{(2)}, \mathbf{z}_{j(n)}^{(2)}) \\ & \left\|\mathbf{x}_n - f\left(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)} \mid \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}\right)\right\|^2 d\boldsymbol{\zeta}^{(1)}\, d\boldsymbol{\zeta}^{(2)} \end{aligned} \quad (3)$$

さらに f が滑らかな写像なので，被積分関数の極値付近での 2 次 Taylor 展開まで考えれば，目的関数は次式で近似できる．

$$\tilde{F}\left[\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}\right] := -\frac{\beta}{2}\sum_{n=1}^{N}\left\|\mathbf{x}_n - f\left(\mathbf{z}_{i(n)}^{(1)}, \mathbf{z}_{j(n)}^{(2)} \mid \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}\right)\right\|^2 \quad (4)$$

4. 提案手法

4.1 勾配法による潜在変数推定

(4) で与えられる近似目的関数 $\tilde{F}$ を最大化する $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$ を求めることが本提案手法の目的である．この方法では非線形写像 f を明示的に計算する必要はない．基底関数等のパラメトリックな表現を必要としないため，潜在空間の次元が増えても計算コストは変わらない．

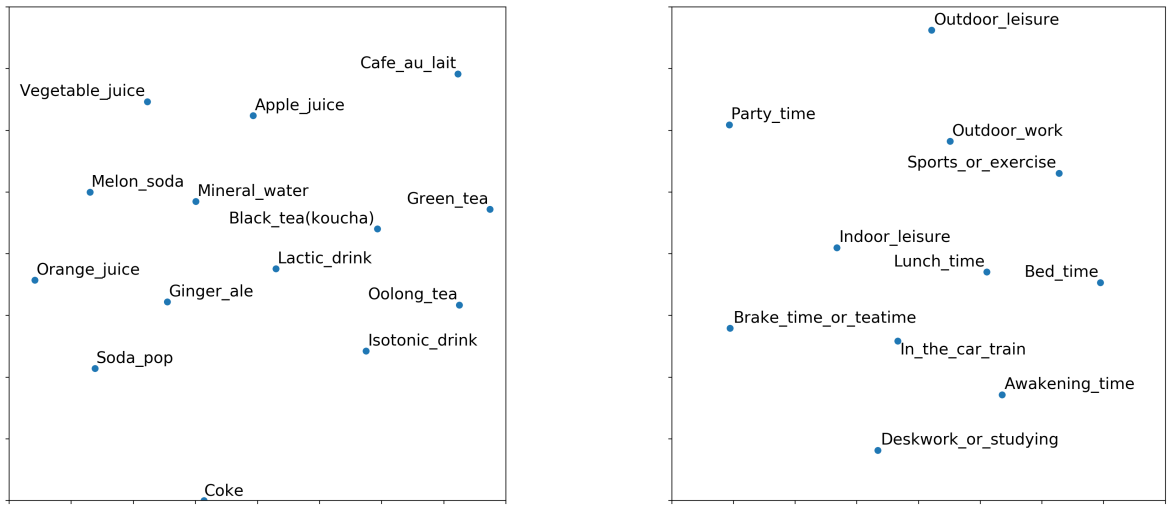
(4) の潜在変数による微分は次式になる．

$$\frac{\partial\tilde{F}}{\partial\mathbf{z}_i^{(m)}} = \beta\sum_{n\in\mathcal{N}_i^{(m)}}\sum_{n'=1}^{N}\left[R_{nn'}\mathbf{d}_{nn}^{\mathrm{T}}\mathbf{d}_{nn'} + R_{n'n}\mathbf{d}_{n'n'}^{\mathrm{T}}\mathbf{d}_{n'n}\right]\boldsymbol{\delta}_{i(n')i(n)}^{(m)} \quad (5)$$

ここで $\mathcal{N}_i^{(m)}$ は同じオブジェクト i から観測されたデータのインデックス集合である．また

$$R_{nn'} := \frac{k(\mathbf{z}_{i(n)}^{(1)}, \mathbf{z}_{i(n')}^{(1)})\, k(\mathbf{z}_{j(n)}^{(2)}, \mathbf{z}_{j(n')}^{(2)})}{\sum_{n''} k(\mathbf{z}_{i(n)}^{(1)}, \mathbf{z}_{i(n'')}^{(1)})\, k(\mathbf{z}_{j(n)}^{(2)}, \mathbf{z}_{j(n'')}^{(2)})}$$
$$\hat{\mathbf{x}}_n \equiv f(\mathbf{z}_{i(n)}^{(1)}, \mathbf{z}_{j(n)}^{(2)}) := \sum_{n'=1}^{N} R_{nn'}\mathbf{x}_{n'}$$
$$\mathbf{d}_{nn'} := \hat{\mathbf{x}}_n - \mathbf{x}_{n'}$$
$$\boldsymbol{\delta}_{ii'}^{(m)} := \mathbf{z}_i^{(m)} - \mathbf{z}_{i'}^{(m)}$$

である．(5) により勾配法を用いて潜在変数を推定するのが提案手法である．



(a) (b)

図 1　Beverage preference data の可視化結果．(a) 飲料の潜在空間（飲料マップ）．(b) 状況の潜在空間（状況マップ）．

4.2　初期化と simulated annealing

目的関数 (1) は非凸関数なので局所解が存在する．そのため局所解を回避するには simulated annealing が有効である．提案手法ではカーネル平滑化を用いているため，カーネル関数 $k(\mathbf{z},\mathbf{z}') = \exp\left[-\|\mathbf{z}-\mathbf{z}'\|^2/2\sigma^2\right]$ の距離定数 σ を初期状態で大きく取っておき，徐々に小さくすることで simulated annealing と同じ効果を得ることができる．これは自己組織化マップ (Self-Organizing Map: SOM) における近傍関数のスケジューリングと同じ意味を持つ [3].

Simulated annealing の効果は距離定数 σ と潜在変数の分布の広さとの相対関係で決まる．すなわち σ を大きく取ることと，潜在変数を小さく分布させることは同じ効果を持つ．そこで距離定数 σ を変える代わりに潜在変数の分布を徐々に広げることで同様の効果が得られる．すなわち初期状態で潜在変数を原点近く（距離定数 σ よりも小さい半径内に）分布させておく．潜在変数の分布は勾配法で更新することで自然と広がっていき，最終的には事前分布に近くなる．したがって σ のスケジューリングをすることなしに自動的に simulated annealing が実現できる．

4.3　可　視　化

関係データの解析には，ドメイン内のオブジェクト関係とドメイン間のオブジェクト関係の両方を可視化する必要がある．本提案法では以下のようにドメイン内・ドメイン間の可視化が可能である．

まずドメイン内可視化については，それぞれのドメインに属するオブジェクトを潜在空間内に写像することで可視化される．たとえば商品評価データの場合，ユーザー集合はユーザー潜在空間へと写像される．ユーザー潜在空間上で類似する嗜好のユーザー $\omega_i^{(1)}, \omega_{i'}^{(1)}$ は似た潜在変数 $\mathbf{z}_i^{(1)}, \mathbf{z}_{i'}^{(1)}$ へ写像される．したがってユーザー空間はユーザーの類似度を表現したマップとして読むことができる．同様に商品集合も商品潜在空間へと写像され，類似した顧客層を持つ商品は近くに配置される．

一方，ドメイン間可視化は写像 $f(\boldsymbol{\zeta}^{(1)},\boldsymbol{\zeta}^{(2)})$ を潜在空間上に表示することで可視化できる．今，ユーザー $\omega_i^{(1)}$ がどのような商品を好むかを可視化したいとしよう．このとき，商品潜在空間を $f(\mathbf{z}_i^{(1)},\boldsymbol{\zeta}^{(2)})$ に応じてグレースケール等で色づけることで，嗜好する商品が存在する領域を可視化することができる．同様に商品 $\omega_j^{(2)}$ を好むユーザー層を知りたいときは，$f(\boldsymbol{\zeta}^{(1)},\mathbf{z}_j^{(2)})$ を使ってユーザー潜在変数を色づければ良い．これは計算機上でインタラクティブに実装できるため，ドメイン間解析を容易に行うことができる．

5.　実　　験

5.1　Beverage Preference Dataset の可視化

本提案手法を用いて Beverage preference dataset [4] の可視化を行った．本データは 604 人のユーザーが 14 種の飲料に関して 11 の状況下で飲む頻度を調査したデータであり，3 次テンソルデータになっている．本実験では飲料と状況の 2 項関係データとみなしてモデル化・可視化を行った．

図 1 に可視化結果を示す．飲料マップ（図 1 (a) では類似する状況で好まれる飲料を表現しており，烏龍茶 (Oolong-tea) と緑茶 (Green-tea)，サイダー (Soda-pop) とコーラ (Coke) などが近くに写像されている．同様に状況マップ（図 1 (b)）では類似した飲料が好まれる状況を表現している．

6.　議論: 他手法との関連

Tensor SOM [5] は SOM を関係データに拡張し，非線形テンソル解析を実現したものである．Tensor SOM と本提案手法は

同じ目的関数 (1) から導出される．Tensor SOM と本提案手法とは以下の点で異なる．

- Tensor SOM は潜在空間を離散化する必要がある．そのため潜在変数も写像も離散点についてのみ求める（ただし基底関数を用いた連続表現版も提案されている [5]）．
- Tensor SOM では潜在空間を離散化するため，潜在空間に関して強い制約を受ける．たとえば $\mathbb{R}^z$ 全体にわたる非コンパクトな潜在空間を考えることはできない．また潜在空間の次元が大きくなると指数的に離散点も増えて計算が困難になる．本提案法はこのような制約を受けず，非コンパクトな潜在空間や次元の高い潜在空間でも計算量を増やすことなく計算できる．また潜在変数の事前分布も一様分布に限定せず自由に決めることができる．
- Tensor SOM では潜在変数と写像を交互推定する．これは広義の EM アルゴリズムとみることができる．一方，提案手法では写像を潜在変数によって決まるものとして目的関数から消去し，潜在変数のみの最適化問題に帰着している．
- 関係データがすべてのオブジェクト組に対して網羅的に与えられる場合，関係データはテンソルデータとみなすことができる．このとき Tensor SOM では一次モデル（インスタンス多様体）を求めることで計算量を大きく減らすことができる．しかしデータが網羅的に与えられない場合（テンソルデータとみなすと欠損が生じる場合）は一次モデルが使えず，計算速度の優位性が失われる．本提案法ではデータが非網羅的であることを前提に作られていて，計算量は変わらない．
- 一方で本手法はノンパラメトリックな表現を用いるため，データ数 N に対して計算量が $O(N^2)$ で増加する．関係データではオブジェクト数 I, J の双方に比例して N が増えるため，中規模データであっても N が大きくなりやすい．この点を改善することが提案手法では必要になる．

7. ま と め

本稿では関係データの可視化法を提案した．本手法は関係データを行列・テンソルとして扱う必要がないため，非網羅的データに対しても応用可能である．またノンパラメトリック表現を用いるため，交互推定を必要とせず，また潜在空間のとり方にも制約を受けない．今後は本手法のモデリング精度の検証および計算コストの削減等を取り組む計画である．

8. 謝　　辞

本研究は JSPS 科研費（課題番号 18K11472）の助成を受けました．また本研究は ZOZO テクノロジーズより助成を受けました．

文　　献

[1] 石黒勝彦, 林浩平,「関係データ学習」, 機械学習プロフェッショナルシリーズ, 講談社, 2016.
[2] Kolda, T.G., Bader, B.W., “Tensor decompositions and applications”, SIAM Rev., vol.51, 2009.
[3] Kohonen, T., “Self-organizing maps”, 3rd ed., Springer, 2001.
[4] http://www.brain.kyutech.ac.jp/~furukawa/beverage-e/
[5] Iwasaki, T., Furukawa, T., “Tensor SOM and Tensor GTM: Nonlinear tensor analysis by topographic mappings”, Neural Networks, vol.77, 2016.