機械学習における関係データのモデル化

鹿島　久嗣

ソーシャルネットワークサービスの普及などによって，従来のデータ解析が対象としてきた表形式とは異なるネットワーク構造をもつデータを解析する需要が高まっている．機械学習やデータマイニング分野においてはネットワーク構造をもつデータはオブジェクト間の関係を表す関係データとして扱われる．本稿ではこの分野でよく利用される行列や多次元配列の低ランク分解に基づくアプローチを中心に解説するとともに，確率的ブロックモデルなどの確率的な関係データの生成モデルについても紹介する．

キーワード：ネットワーク，行列分解，テンソル分解，確率的ブロックモデル

1. 関係データの機械学習

IT 技術や計測技術の進歩によって生み出される膨大な量の情報に立ち向かうべく，学術分野やビジネスなどあらゆる領域において機械学習・データマイニング・統計科学といったデータ解析技術が注目を浴びており，ビッグデータの旗印の下で，その研究，開発，そして応用が進んでいる．多くの場合，データ解析の興味の対象は個々のオブジェクト（人・物・事）の性質であり，たとえば，ある患者が特定の疾患をもつかとか，ある顧客が特定のキャンペーンに対し興味を示すか否か，あるいはある薬剤候補化合物が特定の性質をもっているか否かといったことなどを，データから導き出すことを目的とする．一方，個々のオブジェクトではなく，複数のオブジェクトの間の「関係」の解析に対する興味が高まっている（図 1）．たとえばソーシャルネットワーク分析においてはある領域で活動する人間同士の関係が，オンラインショッピングサイトにおけるマーケティングではそこを訪れる多数の顧客とさまざまな商品の間の関係がその興味の対象となる．また，生命科学や創薬の分野においては，種々のタンパク質の間の，あるいはタンパク質と化合物との間の相互作用が形作る生体ネットワークが重要な役割を担うが，これらもタンパク質や化合物の間の関係を表したものである．このようにさまざまな領域においていわゆる“関係データ”が注目されるようになり，結果としてこれら関係データを対象とした分析が盛んになりつつある．

関係データとは複数のオブジェクトの組についてのデータと位置づけることができる．その解析においては，関係の成立や関係のもつ性質についての推論を行うことが多く，たとえばオンラインショッピングサイトなどにおける推薦システムでは顧客と商品との間の関係（購買，評価，情報の閲覧など）を予測し顧客に適切な商品の推薦を行う．あるいはソーシャルネットワークサービスにおいては友人やコミュニティの推薦機能が提供されているが，これらもユーザーやコミュニティといったオブジェクトの間に成り立つ関係の推論に基づくものである．製薬会社が新規薬剤候補の効率的な探索を行うためには，薬剤候補となる化合物とその標的となるタンパク質の間の相互作用の予測を行うことで薬剤候補のスクリーニングを行うことが重要である．

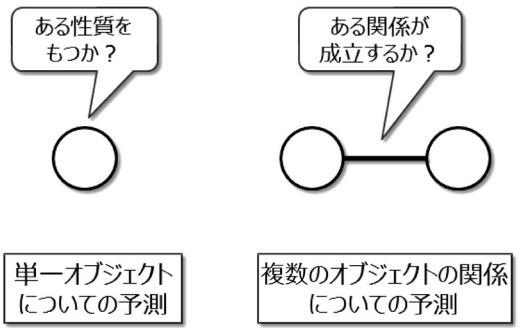

図 1　通常のデータ解析（左）と関係データ解析（右）

本稿は，これら関係データを対象とした解析を行う際に用いられるモデルや手法，そしてその近年の発展について，主に機械学習・データマイニングの立場から紹介するものである．当然ながら関係データ解析の技術は必ずしもこれらの分野でのみ発展してきたものではなく，伝統的には統計科学や社会科学，あるいはオペレーションズ・リサーチにおいても扱われてきたものである．しかしながら前述のビッグデータの潮流

かしま　ひさし
京都大学
〒 606–8501 京都府京都市左京区吉田本町 36–1

などを背景として，近年，機械学習・データマイニング分野で大きな注目を浴び発展している．そこで，以下ではまず，2 オブジェクト間の関係データの基本的なモデルとして頻繁に用いられる低ランク行列による解析について述べた後，その多オブジェクト間関係への拡張であるテンソル分解を紹介し，機械学習・データマイニング分野におけるこれらの研究動向について述べる．後半では，関係データの確率的な生成モデルとして確率的ブロックモデルについて紹介する．

2. 低ランク行列による関係データのモデリング

まずは最も単純な場合として二つのオブジェクト間に成立する一種類のみの関係のモデルを考える．このような場合の関係データの表現としてはしばしば行列が用いられる．これはオブジェクトを頂点として，また二つのオブジェクトの間にある関係が成り立つ場合に辺を置くことでデータ全体をグラフとして表現した場合の隣接行列に相当する．つまり，i 番目と j 番目のインデックスをそれぞれもつ二つのオブジェクトの間に関係が成り立つならば，対称なデータ行列 $\boldsymbol{X}$ の (i,j) 要素 $x_{i,j}$（および $x_{j,i}$）は値 1 をとり，そうでないならば 0 をとる．

関係データ解析において，しばしば $\boldsymbol{X}$ が近似的に低ランク行列であるという仮定を置く．つまり，$\boldsymbol{X}$ が概ね $\boldsymbol{X} \sim \boldsymbol{U}\boldsymbol{U}^\top$ と表現できるとみなす（図 2）．ここで「$\sim$」と書いたのは近似的に等号が成り立つということであり，多少の雑音成分を無視すれば概ね $\boldsymbol{X}$ は $\boldsymbol{X}$ よりも“薄い”行列 $\boldsymbol{U}$ の行列積で書けるということである．$\boldsymbol{U}$ の厚さが K，すなわち $\boldsymbol{X}$ のランクが K の場合，$\boldsymbol{U}$ の各行は各オブジェクトを K 次元のベクトルで表現したもの（言い換えれば K 次元の“潜在的”空間において各オブジェクトが配置されているもの）と解釈することができる（図 3）．なお，上記のモデルは関係に参加するオブジェクトに順序がない場合，たとえば友人関係のような場合（つまり，i と j が友人関係にあることと j と i が友人関係にあることが同じ意味となるような場合）を想定しているが，オブジェクトに順序がある場合，たとえば「一方から他方に対してメッセージを送った」というような関係の場合には，行列分解モデルは二つの行列 $\boldsymbol{U}$ と $\boldsymbol{V}$ を用いて $\boldsymbol{X} \sim \boldsymbol{U}\boldsymbol{V}^\top$ という分解を考える．

観測によって得られたデータ行列に対して行列分解を行い $\boldsymbol{U}$（と $\boldsymbol{V}$）を推定する場合，上で「$\sim$」で表した行列間の近さを具体的な誤差関数として定義し，これを最小化するような $\boldsymbol{U}$（と $\boldsymbol{V}$）を求める最適化問題を解くことになる．これは与えられたデータ行列 $\boldsymbol{X}$ をもっともよく近似する低ランク行列 $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{U}^\top$ を求めることに相当する．たとえば誤差を 2 乗誤差（行列の差のフロベニウスノルム）で測り，ランクを K の場合に限定するという制約を置くと，最適化問題は

$$\text{minimize} \ \| \boldsymbol{X} - \boldsymbol{Y} \|_F^2 \ \text{s.t.} \ \text{rank}(\boldsymbol{Y}) \leq K$$

となる．これらの最適化問題は通常凸最適化問題にはならないが，上記の場合には最適解は特異値分解によって得ることができる [1].

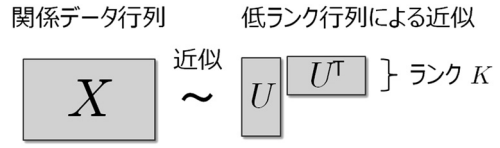

図 2 低ランク行列による関係データの近似

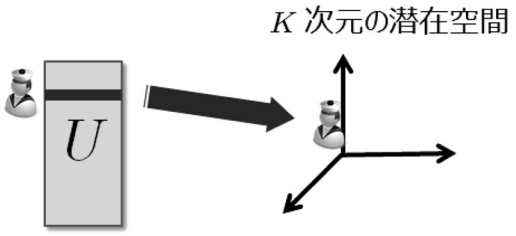

図 3 K 次元の潜在空間におけるオブジェクトの表現

オンラインショッピングサイトやソーシャルネットワークを対象とした場合，扱うオブジェクト数が膨大になるため，大きな最適化問題を解くための工夫が必要となる．たとえば，データ行列全体を使って一度に最適化を行うのではなく，データの一部（たとえば一要素 $x_{i,j}$）のみを用いた解の更新を繰り返すといったアプローチもとられる．また，上で述べたように多くの場合，解くべき最適化問題は凸最適化問題とならない [1]．これは行列の低ランク制約は凸集合を作らないためである．この問題を解決するために，凸集合となるトレースノルム制約が低ランク制約の代わりに用いられる [1]．トレースノルムとは行列の特異値の和であり，これに制約を加えるということは間接的にランクに制約を与えていることになる．

3. テンソル分解による複雑な関係のモデル化

前節では二つのオブジェクト（たとえば 2 人の人間）の間に成立しうる単一種類の関係（友人関係）を対象

としたモデル化を低ランク行列により行ったが，世の中にある関係データは行列として表現できるものにとどまらない．たとえば三つ以上のオブジェクトの間の関係や複数種類の関係，あるいは関係の成立条件などがあるような場合もあるだろう．自然言語処理を例にとれば，主語・述語・目的語の関係は三つのオブジェクトの間に成立する関係となる．別の例としてオンラインショッピングサイトを考えてみると，客が商品に対してとることのできる行動は「購入」の一種類だけではない．他にも「商品情報の閲覧」や「商品をショッピングカートに加える」さらには「評点の入力」などさまざまな種類の行動がありうる．あるいは，ある特定の時点で起こる関係や時間的に変化する関係など，時間的要素は関係の成立条件として典型的なものである．

三つ以上のオブジェクトの関係は，行列の一般化である多次元配列として表現することができる．（関係の種類・時間要素などもある種のオブジェクトとして考えることで，三つ以上のオブジェクト関係として統一的にとらえることができる．）多次元配列のモデルとしてしばしば用いられるのが，行列の低ランク分解を多次元配列に拡張したテンソル分解 [2] である．テンソル分解では，D 次の多次元配列（D オブジェクト間の関係）をコアテンソルと呼ばれる小さな多次元配列と D 個の因子行列に分解する（図 4）．これはちょうど 2 オブジェクト間関係を対象としたときには ($D = 2$)，コアテンソルを単位行列にとることで前述の行列分解に一致するという意味で，行列分解の一般化となっている．テンソル分解にはいくつかの種類があるが，中でも代表的なものが CANDECOMP/PARAFAC (CP) 分解と Tucker 分解である．CP 分解は行列の特異値分解をテンソルに拡張したものであり，そのコアテンソルは対角成分のみが非零の値をもつのに対し，Tucker 分解は密なコアテンソルをもつ．

テンソル分解を求めるときにも行列分解の場合と同様，与えられた関係データ多次元配列を低ランクテンソルで近似する最適化問題を解くことなる．多次元配列の次数が大きくなるとモデルが複雑になるため解くべき最適化問題も複雑になるが，より大規模な多次元配列を効率的に扱うためのアルゴリズムが精力的に研究されている．

4. テンソル分解による関係データモデリングの発展

予測はデータ解析の中でももっとも重要なタスクのうちの一つであり，これは関係データ解析においても

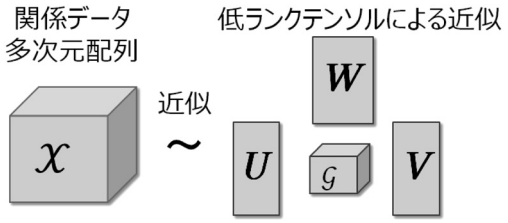

図 4　低ランクテンソル分解による関係データの近似

同様である．これまでに例として挙げてきたオンラインショッピングサイトやソーシャルネットワークサービスなどにおける推薦も，予測に基づいたものである．多次元配列によって表現された関係データの予測問題は多次元配列の補完問題として定式化される．これは，多次元配列の要素が部分的に与えられたときに，これらの観測要素を手掛かりに未観測の要素を予測するような問題である．この補完問題をテンソル分解によって解く場合，まずは多次元配列を低ランクテンソルで近似する最適化問題を解く．そして得られた低ランクテンソルを掛け合わせ元の多次元配列を復元することで，観測値では零だが（つまり関係がないが）復元値は 1 に近い値をとる箇所が現れる．これらが将来関係が発生すると期待できる箇所と解釈できる．

現実の関係データでは，多次元配列中において観測される部分が全要素数に比較してずっと少ないということがしばしば起こり，そのうえほとんどのオブジェクトは少数の関係にのみ関与する，すなわち非常に偏りの大きなデータとなることが指摘されている．このようなデータの疎性はモデルの精度の著しい低下をもたらし，その結果，将来の商品の購入や友人関係の発生などの予測においても予測性能の低下を引き起こすため，テンソル分解を用いた関係予測における重要な課題として認識されている．

行列分解のところでも述べたように低ランク制約は非凸制約となるため最適化問題の大局解が得られることは保証されないが，これはテンソル分解においても同様であり，最適化問題としての性質はよくない．この問題は観測データが疎である場合に特に顕著になり，局所解の影響によって時に著しく予測性能が悪化し，またその結果も極めて初期値などのパラメータ依存性の高い不安定なものとなってしまう．この問題に対処するために，近年では大局解が得られるようなテンソル分解の定式化を考えるという流れがある（たとえば [3]）．行列分解を凸最適化問題として定式化する際に用いたトレースノルム制約の考え方をテンソル分解にも拡張し，多次元配列の行列展開（3 階の多次元配

列であれば3種類の展開が考えられるため三つの行列が得られる）のすべてに対してトレースノルム制約を課すことによってテンソル分解を凸最適化問題として定式化することができる．

別の方向性としては，モデルの単純化がある．これは，D 個のオブジェクトの関係を D 個のオブジェクトすべての絡み合いによって表現するのではなく，任意の2オブジェクトの関係の積み重ねによって表現するというものである [4]．たとえば，客と商品そして行動の3オブジェクト間の関係の強さを，客と商品の間の関係，商品と行動との関係，客と行動との関係の3組の関係の強さの和として予測するのである．現実世界の多くの関係においては D 個すべてのオブジェクトがあって初めて成立するような関係は稀であるため，通常のテンソル分解ではモデルが過剰に複雑である場合もあり，このような単純化によって問題が解きやすく，また実データにおける予測性能も良好であることが確認されている．これとよく似たモデルの単純化を用い，固有値問題を1回だけ解くことによって大局解が得られるような定式化も提案されている [5]．

上記のアプローチはテンソル分解の定式化を改善あるいは簡略化することでデータ疎性に対処するというものであったが，新たな情報を付加することによって解決を図るというアプローチもある．多くの場合において関係データ以外にも外部情報が存在するため，これらの有効活用もまた必須である．たとえば，顧客と商品の購買関係を考えた場合，購買情報の他にも，顧客の個人情報や商品の情報などが利用可能である．このような場合には，互いによく似た顧客，よく似た商品が似た振る舞いをすると仮定するのは自然であろう．これらの事前知識を行列分解あるいはテンソル分解の最適化問題に明示的に制約として取り入れるというアプローチも提案されている [6, 7]．結果として，類似するオブジェクト同志で，テンソル分解によって得られる因子行列の各オブジェクトに対応する部分が類似するようになる．このように補助情報を利用することで観測データが極めて疎な場合であっても予測性能の低下を緩和する効果があることが示されている [7]．

5. 関係データの確率モデル

前節までで紹介した行列・テンソルの低ランク分解に基づくモデルは，既存の数値計算パッケージや効率のよいアルゴリズムなどを利用した大規模計算が可能であったり，定式化によっては理論的に大域解が保証されているなどという点において，関係ビッグデータの処理には適したものである．しかしこれを逆に考えると，ある意味“計算の都合”に縛られたモデル化となっており，われわれがデータにどのような仮定を置いているか，すなわち対象に対するわれわれの事前知識が必ずしも反映されたものではないという可能性もある．これに対し，データがどのように確率的に生成されるかという過程を確率モデルとして表現する生成モデルのアプローチもとられている[1]．生成モデル化のイメージとしては，たとえば万物を創造する創造主のような存在がいたとして，その創造主が（サイコロを振りながら）どのようにデータを生成したかという観点からデータをとらえていると考えるとわかりやすい．万能の創造主と言えども，万物のそれぞれを（さらに関係データの場合，それらの間の関係を）いちいちデザインするのは大変なので，これらを生成するための確率的なルールだけを定めておき，あとは乱数を使って自動的にこれらが生成されるようにするのである．

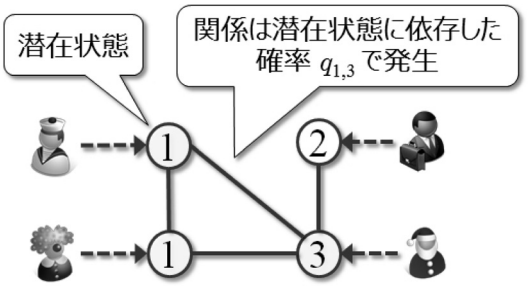

図5 確率的ブロックモデルにおける関係の生成

関係データの確率モデルとしてもっとも代表的なものの一つが確率的ブロックモデル [9] である．古くより社会ネットワーク分析などにも用いられるこのモデルは，各オブジェクトがそれぞれ潜在的な変数をもち，その潜在的変数の組み合わせによって二つのオブジェクト間に関係が発生するというモデルである．確率的ブロックモデルでは，次のようなデータ生成過程を考える．まず，各オブジェクトがとりうる潜在的な状態が K 通りあるとする．各オブジェクトは K 通りの可能性のうえでの離散的な確率分布によって K 通りのうちの一つの状態に決まるものとする．より具体的に言えば，オブジェクト i は，それぞれその状態を決める確率変数 S_i をもつとし，S_i が状態 k に決まる確率が $\Pr(S_i = k) = p_k$ であるということである．そのうえで，各オブジェクトの組 (i, j) の間に関係が発生するかどうかがベルヌーイ分布によって決定されるとす

[1] ただし，行列分解の確率的な生成モデルとしての解釈もある [8]．

る．このベルヌーイ分布は二つのオブジェクトの状態の組ごとに決まり，二つのオブジェクトの状態がそれぞれ k, ℓ であった場合，確率 $q_{k,\ell}$ で関係が発生する（図 5）．

確率的ブロックモデルの関係データ生成過程をソーシャルネットワークを例に説明するならば，世の中に K 種類のタイプの人間が存在し，それぞれのタイプの人間の割合が $(p_1, p_2, \ldots, p_K)$ で決まっていると考える．たとえば，タイプが性格を表すとすると「おっとりしている」とか「積極的である」とかの性格のタイプが K 通りあるという感じである．前述した創造主の視点から，それぞれの人間はまずこの分布に従って K 通りのタイプのいずれかに振り分けられる．そして次に，K タイプの人間の任意の二つのタイプの組み合わせに対して，その間の相性のようなもの $((q_{k,\ell})_{k,\ell=1,2,\ldots,K})$ を定め，これによって友人関係が定まるといった具合である．たとえば「おっとりしている」同士，「積極的である」同士だと意外に友人関係にはなりにくいが，「おっとりしている」人と「積極的である」人は意外に馬が合うといった感じである．

さて，実際の人間関係を考えてみると，たとえば職場における人間関係とプライベートにおけるそれは異なるといったように，関係ごとにその文脈が異なる場合というのは容易に想像できる．確率的ブロックモデルを発展させた混合メンバシップ確率的ブロックモデル [10] と呼ばれるモデルでは，関係を生成するごとに潜在状態を振りなおすことで関係ごとに生成される文脈を考慮する．

確率的ブロックモデルのパラメータと各オブジェクトの潜在状態の推定には，最尤推定やベイズ推定が用いられる．その計算は前節までのモデルと比較すると容易ではなく，扱うことのできる状態数やオブジェクト数は限定され，実際の計算には近似計算やマルコフ連鎖モンテカルロ法などが用いられる．

6. おわりに

本稿では機械学習・データマイニングの分野でよく利用される行列や多次元配列の低ランク分解に基づくアプローチを中心に，関係の確率的な生成モデルである（混合メンバシップ）確率的ブロックモデルについても紹介した．関係データのモデル化については古くから興味をもたれてはきたが，ソーシャルネットワーキングサービスや推薦システムといった重要性の高い応用と大規模データの出現を契機に近年大きく研究や実用が進んでいるテーマであり，研究や応用の方向性も多岐にわたる．本稿がそのすべての動向をカバーできているとは決して言えないが，関係データモデリングの基本的な考え方を伝えることができたなら幸いである．

参考文献

[1] M. Fazel, “Matrix rank minimization with applications,” Ph.D. thesis, Stanford University, 2002.

[2] T. Kolda and B. Bader, “Tensor decompositions and applications,” *SIAM Review*, **51**, pp. 455–500, 2009.

[3] R. Tomioka, T. Suzuki, K. Hayashi and H. Kashima, “Statistical performance of convex tensor decomposition,” *Advances in Neural Information Processing Systems*, **24**, pp. 972–980, 2011.

[4] S. Rendle and L. Schmidt-Thieme, “Pairwise interaction tensor factorization for personalized tag recommendation,” In *Proceedings of the 2010 ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 81–90, 2010.

[5] N. Nori, D. Bollegala and H. Kashima, “Multinomial relation prediction in social data: A dimension reduction approach,” In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 115–121, 2012.

[6] W. J. Li and D. Y. Yeung, “Relation regularized matrix factorization,” In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1126–1131, 2009.

[7] A. Narita, K. Hayashi, R. Tomioka and H. Kashima, “Tensor factorization using auxiliary information,” *Data Mining and Knowledge Discovery*, **25**, pp. 298–324, 2012.

[8] A. Mnih and R. Salakhutdinov, “Probabilistic matrix factorization,” *Advances in Neural Information Processing Systems*, **20**, pp. 1257–1264, 2007.

[9] T. A. Snijders and K. Nowicki, “Estimation and prediction for stochastic blockmodels for graphs with latent block structure,” *Journal of Classification*, **14**, pp. 75–100, 1997.

[10] E. M. Airoldi, D. M. Blei, S. E. Fienberg and E. P. Xing, “Mixed membership stochastic blockmodels,” *Journal of Machine Learning Research*, **9**, pp. 1981–2014, 2008.