

解 説

関係データの機械学習

—行列・テンソル分解によるアプローチ—

Machine Learning for Relational Data
— Matrix and Tensor Factorization Approaches —

鹿 島 久 嗣* *京都大学大学院情報学研究科知能情報学専攻
Hisashi Kashima* *Kyoto University

1. は じ め に

近年の計測技術の目覚ましい進歩は様々な対象を計算機上で取り扱い可能なデータとして記述することを可能にした。そして20世紀の終わりごろから始まったいわゆる情報技術革命によって様々な企業や自治体で情報インフラが導入され、以降様々なデータが組織の中で蓄積されるようになった。一方、それ自身も巨大なデータの塊として見ることもできるウェブの出現は、組織や地域を超えたデータの流通を促した。こうして世の中の様々な事象が大量のデータとして蓄積・記述され、アクセス可能になると、今度はいかにこれらを分析して価値を生み出していくかということに人々の関心が移っていく。最近のビッグデータの潮流はまさにこの変化を表しているといえる。データを中心とした考え方は、ビジネス・科学を含むあらゆる分野に浸透し始めている。そして、機械学習・データマイニング・統計科学などのデータ解析技術は、ビッグデータ分析を行うための基盤技術として大きな注目と期待を集めている。

さて、データ解析の手法は数多あるが、これらの主な興味は個々のオブジェクトのもつ性質、例えばマーケティングの文脈では、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ということであったり、あるいは製薬会社であれば、ある薬剤候補となる化合物が所望の性質をもっているか否かなどにある。そして近年、その興味は個々のオブジェクトだけでなく、それらの間の「関係」へと拡大しつつある。例えば、ソーシャルメディアやオンラインショッピングの普及を背景として、SNSでは参加者同士の関係の、オンラインショッピングサイトでは顧客と商品の間の関係の解析に基づく推薦機能が活躍している。生命科学や創薬の分野では、タンパク質同士、あるいは化合物とタンパク質の間の作用の解析が盛んに行われている。これらのデータはオブジェクト間の関係を表す「関係データ」と呼ばれ、

その解析手法の研究はデータ解析分野における大きなトレンドの一つとなっている。

関係データは複数のオブジェクトの組についてのデータ、すなわちある種のメタデータと位置付けることができる。多くの場合、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を目的とする。例えば上で例として挙げた推薦システムでは、オンラインショッピングサイトにおける顧客と商品との間の関係（多くの場合、購買関係が主な興味となる）を予測し顧客に適切な商品の推薦を行ったり、SNSにおいて友人やコミュニティなど新たな繋がりや提案を行ったりという目的で用いられる。あるいは、薬剤候補となる膨大な数の化合物とその標的となる多くのタンパク質の間の相互作用の予測を大規模に行うことで有望な組を発見できれば、製薬会社の創薬プロセスを効率化することができるだろう。

本稿では、このような関係データの解析手法について、特に行列分解・テンソル分解を用いたアプローチの解説を行う。まず2オブジェクト間の関係データの表現である行列を対象とした行列分解アプローチを紹介する。次に、より一般的な関係の表現として、三つ以上のオブジェクトの関係を表現できる多次元配列を取り上げ、その解析方法としてテンソル分解によるアプローチを紹介する。最後に実際の関係データの解析を行う際に直面する様々な課題に対処するための、テンソル分解の様々な発展について紹介する。

2. 行列による関係データのモデル化

2.1 モチベーション：推薦システム

具体的な題材として、関係データ解析の主要な適用先である推薦システムを取り上げ話を進めることにする。推薦システムとは、個人に適応した情報推薦の仕組みであり、その目的は個人の利用者のそれぞれにカスタマイズされたきめ細かいサービスの実現にある。現在では推薦システムは、ショッピングサイトをはじめとして、SNSの友人推薦やニュース記事の推薦など、ウェブ上の様々なサービスで提供されている機能である。

推薦システムは機械学習のキラーアプリケーションの一つとして盛んに研究されているが、ここ数年の推薦技術の

原稿受付 2014年12月3日

キーワード: Machine Learning, Relational Data, Matrix Factorization, Tensor Factorization

*〒606-8501 京都市左京区吉田本町

*Sakyo-ku, Kyoto-shi, Kyoto

	商品			
	車	携帯電話	シャツ	ヘルメット
ユーザ1	1	0	1	0
ユーザ2	0	1	1	0
ユーザ3	0	1	0	1

ユーザ

商品

購買関係

図1 行列による関係データの表現

発展のきっかけを作ったのが2006年から2009年までの間、米国のオンラインビデオレンタル会社であるNetflix社によって開催されたNetflix Prizeである。同社の推薦アルゴリズムの予測性能を改善することを目的として100万米ドルの賞金を懸けて行われたコンテストは大きな話題を呼び、研究者からエンジニアまで様々な人々が時にチームを組んで取り組んだ。その結果として推薦アルゴリズムの技術開発が大きく進展することとなった。このコンテストで好成績を収めた手法が用いていたのが行列分解によるアプローチである[1]。

2.2 行列による推薦システムのモデル化

オンラインショッピングサイトにおける商品推薦問題は、ユーザの集合と商品の集合の間に成り立つ購買関係の予測問題として捉えることができる。あるユーザがある商品を購入するか、しないかという行動を0,1の2値変数によって表現するとしよう。すべてのユーザとすべての商品の間の購買関係は行列 $\mathbf{X}$ としてまとめて書くことができる(図1)。 $\mathbf{X}$ の (i,j) 要素 $x_{i,j}$ が1のときは i 番めのユーザが j 番めの商品を買ったことを表し、0のときは買わなかったことを表す。ここで、現在は0であるが、近い将来あるいは潜在的には購買関係が成り立つはず(つまり1であるはず)であるようなユーザと商品の組を見つけることができれば、そのユーザにその商品を推薦すればよいということになる。

なお、推薦問題の別の定式化としては、ユーザが商品に対して評点を与えるような状況において、評点を与えていない組(すなわち未観測値)に対してその評点を予測するという定式化もあり、推薦システムの問題設定としてこちらのほうが一般的ではあるが、本稿は関係予測の観点からの解説であるため、前述の設定に基づくことにする。ただし、とりうるアプローチとしては本質的にはそれほど変わらない。

2.3 行列分解による関係予測

行列分解によるアプローチでは、与えられた関係行列 $\mathbf{X}$ を、 $\mathbf{X}$ より小さな行列の積で近似するというを行う。より具体的には $\mathbf{X}$ が $M \times N$ 行列である(つまり M 人のユーザと N 種類の商品がある)とすると、 $M \times K$ 行列 $\mathbf{U}$ と $N \times K$ 行列 $\mathbf{V}$ (なお、 K は M と N のいずれ

$$\text{ユーザ} \quad \text{商品} \quad \mathbf{X} = \mathbf{U} \mathbf{V}^T \quad \text{ランク } K$$

図2 行列分解

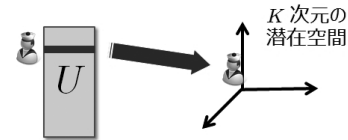


図3 行列分解の解釈

よりも小さい正の整数とする)を用いて

$$\mathbf{X} \sim \mathbf{U} \mathbf{V}^T \quad (1)$$

のように近似を行う(図2)(ここで「 $\sim$ 」はその両辺が何らかの意味で近いことを表すものとする)。つまり、関係式(2)は $\mathbf{X}$ が多少の雑音を取り除けば本質的には低ランク(上記 K がランクに対応)であることを仮定していることを意味している。

$\mathbf{U}$ と $\mathbf{V}$ をどのようにして得るかは後ほど述べることで、いったん上記のような $\mathbf{U}$ と $\mathbf{V}$ が求まったものとして、これらを用いてどのように予測を行うかを述べる。上記の関係式(1)を要素ごとに書けば

$$x_{i,j} \sim \mathbf{u}^{(i)} \mathbf{v}^{(j)T} \quad (2)$$

となる。ここで $\mathbf{u}^{(i)}$ は $\mathbf{U}$ の i 行め、 $\mathbf{v}^{(j)}$ は $\mathbf{V}$ の j 行めを取り出したものとする。関係式(2)における近似の程度はユーザと商品の組み合わせ (i,j) によって異なり、ほぼ同じ値となるのものもあれば、大きくずれているものもあるだろう。ここで注目すべきは、 $\mathbf{X}$ においては $x_{i,j} = 0$ であるが、その近似値 $\mathbf{u}^{(i)} \mathbf{v}^{(j)T}$ が0より大きくなるようなユーザ(i)と商品(j)の組み合わせである。このような組み合わせに対しては、現在は購買関係が成立していない($x_{i,j} = 0$)が、将来あるいは潜在的には関係が成立する($\mathbf{u}^{(i)} \mathbf{v}^{(j)T} \sim 1$)ものと期待できる。

$\mathbf{u}^{(i)}$ と $\mathbf{v}^{(j)}$ がそれぞれ何を意味するのかを解釈してみよう。 $\mathbf{u}^{(i)}$ と $\mathbf{v}^{(j)}$ はそれぞれ K 次元のベクトルであることに注意すると、これは各ユーザと各商品がそれぞれ K 次元の空間の一点として表現されていることに相当する(図3)。すると、上記の予測法はこの空間において近い(内積が大きい)ユーザと商品の間に関係が成り立つと予測することになる。また、同じような嗜好を持つユーザや、同じような特徴をもつような商品はこの空間内で近くに配置される。この空間は購買関係という観点から、ユーザと商品を K 軸の要素で表現した潜在的(抽象的)な空間であり、実際にこの空間に配置されたユーザや商品を検証することで、後付けで軸の意味(「娯楽要素」とか「価格」など)の解釈を行うことのできる可能性はあるが、それぞれの軸に

対してあらかじめ明示的な意味が与えられているわけではない。

2.4 行列分解の方法

先ほど先送りしていた、関係式 (1) を満たすような U と V を得る方法について説明する。具体的な手続きとしては、関係式 (1) を達成するという目標を最適化問題として定式化し、これを解くことになるが、そのためにはこれまで定義が曖昧であった「 $\sim$ 」の定義を厳密に行う必要がある。例えば「 $\sim$ 」の意味を各要素の 2 乗誤差の意味に近いということにするのであれば、損失関数 L を

$$L(U, V) = \|X - UV^T\|_F^2 = \sum_{i,j} \left(x_{i,j} - u^{(i)} v^{(j)\top} \right)^2$$

と定義し、これを目的関数として最小化するような U と V を求めることになる（なお $\|\cdot\|_F$ はフロベニウスノルムであり、そのあとの展開式に示されるように行列の各要素の 2 乗和として定義される）。この最小化問題の最適解は、 X の特異値分解を行い、得られた特異値を大きいほうから K 個分のみ残すことで求めることができる。

あるいは本来関係の有無は本来 2 値変数であるため、これを関係の存在確率としてモデル化するのが妥当であると考えられるかもしれない。その場合、確率をロジスティック回帰モデルとして考え、2 乗誤差の代わりに下記の負の対数尤度関数を損失関数として用いるということも考えられる。

$$L(U, V) = - \sum_{i,j} \log \left(\frac{x_{i,j}}{1 + \exp(-u^{(i)} v^{(j)\top})} + \frac{1 - x_{i,j}}{1 + \exp(u^{(i)} v^{(j)\top})} \right)$$

3. テンソル分解による複雑な関係データのモデル化

3.1 多次元配列による関係データの表現

ここまでは二つのオブジェクト間に成立する関係に注目し、これを行列で表現することを考えてきたが、より一般的には三つ以上のオブジェクトの間に成立する関係や、異なる複数の種類の関係、あるいは関係の成立条件等を考えたい場合もあるだろう。例えば、ユーザが商品に対してとることのできる行動は「買う」か「買わない」という一種類だけではなく、「商品情報の閲覧を行う」とか「評価を行う」など様々な種類の行動がありうる。また、ある時点で起こる関係や時間的に状態が変化する関係など、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。そして、おそらくこれらの異なる種類の行動の間には何らかの相関関係があるはずであり、関係データの解析においてこれらを積極的に利用していきたいという場面もあるだろう。このような 3 要素以上の関係の表現として、現在最も良く用

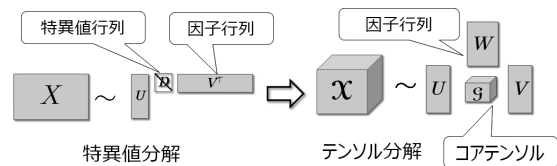


図4 三次元配列のテンソル分解

いられているものが行列の自然な拡張である多次元配列である。したがって関係データの解析は、多次元配列形式をもったデータの解析として捉えることができる。

3.2 テンソル分解による関係予測の課題

行列データの解析にあたり、そのモデルとして行列の低ランク分解を用いたが、これを多次元配列に一般化したものがテンソル分解 [2] である。テンソル分解には、データの性質に対する仮定によって様々な種類のものがあるが、中でも代表的なものとして知られているのがCANDECOMP/PARAFAC (CP) 分解と Tucker 分解である。テンソルの低ランク分解は、 D 次元の多次元配列をコアテンソルと呼ばれる小さな D 次元配列と D 個の因子行列に分解する（図4）。特異値分解が行列すなわち二次元配列を、小さな二次元配列（特異値を対角成分にもつ対角行列）と二つの因子行列に分解するものと考え、テンソル分解はその多次元拡張として見る事ができる。行列の掛け算に対応する演算が多次元配列でどのように定義されるかといった具体的な演算についてはここでは省略するが、CP 分解は行列の特異値分解の素直な拡張になっており、そのコアテンソルは対角成分のみが（特異値に対応する）非零の値を持つのにに対し、Tucker 分解はよりコンパクトな表現になっており、密なコアテンソルを持つ。

テンソル分解を用いて関係予測問題にアプローチする場合、行列の場合と同じように、データとして与えられた多次元配列 $\mathcal{X}$ をうまく近似するようなテンソル分解（コアテンソルと複数の因子行列）を求めることになる。行列の場合と同様に、まずは損失関数を定義してこれを最小化するような最適化問題を解き、得られた分解からもとの多次元配列を再構成することによって予測を得るという一連の手続きは同じである。テンソル分解のアルゴリズムとしては、行列の特異値分解を繰り返し適用する方法等が用いられる [2]。大規模なデータに対してはより単純な勾配法等の最適化手法が使われることも多い。ただし後述するようにテンソル分解の最適化問題は多くの場合非凸最適化問題であるため、これらの解法で得られるのは局所解である。

行列分解で得られた因子行列を解釈したときと同様に、テンソル分解によって得られた因子行列もまた、各オブジェクトを潜在空間上で表現したものと考えられる。やはり、この空間内では似た性質をもったオブジェクトは近くに配置されることになる。

テンソル分解によって、Web ページのタグ推薦（人×Web ページ×タグの三次元配列）や、時間変化するソーシャルネットワークの分析（人×人×時間の三次元配列）、Web のリンク解析（Web ページ×Web ページ×アンカーテキストの三次元配列）、画像認識（画像×人×向き×明るさ×…の多次元配列）など、行列では十分に捉えることができなかったより深い関係データの解析が可能になっている。

4. テンソル分解の発展

4.1 多次元配列の疎性への対処

テンソル分解による関係データ解析は近年機械学習やデータマイニング分野において注目されており、様々なモデルの拡張や、大規模データに対して適用可能なアルゴリズムの開発等が精力的に行われているが、まだ課題も多い。その課題の一つがデータ疎性の問題である。一般にテンソルの観測部分は全要素数に比較してずっと少ないことが多い。例えば多くのソーシャル系サービスにおけるデータでは、オブジェクトのすべての可能な組み合わせに対してその 0.01% 程度の関係しか観測されておらず、さらにほとんどのオブジェクトは少数の関係にのみ関与することが指摘されている。多次元配列の疎性は著しい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。

疎な多次元配列のテンソル分解が難しい理由の一つはその最適化問題としての性質の悪さである。行列分解の定式化のところで見たのと同様、テンソル分解もなんらかのランク制約の下で、与えられた多次元配列を近似する最適化問題として定義されることが多い。ランク制約を満たす解集合は通常は非凸集合となるため、結果として解くべき問題は非凸最適化問題となり、必ずしも大局解が得られるという保証はない。行列分解の場合、2 乗誤差を目的関数とするのであれば特異値分解によって最適解を得ることができたが、一般にテンソル分解においてそのような都合のよいことはない。観測が疎である場合には、問題が特に局所解の影響を受けやすくなり、結果も極めて不安定となってしまう、時に著しく予測性能が悪化する。

このような課題に対する対処法は様々考えるが、一つの方法はモデルを単純化することである。通常の CP 分解や Tucker 分解のように、 D 次元配列を 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単純化を行うモデルが提案されている [3]。現実世界の多くの関係は、必ずしも D 個のオブジェクトがあって初めて成立するようなものだけではないため、このような単純化を行っても実データにおける予測性能は良好であることが確認されている。

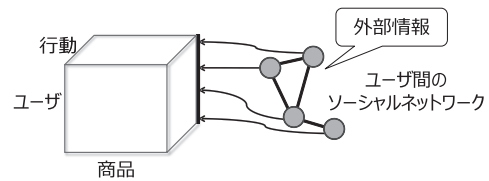


図5 外部情報を利用したテンソル分解

4.2 最適解を保証するテンソル分解

テンソル分解の最適化問題としての性質の悪さを改善するために、近年では大局解が保証されるテンソル分解の定式化を考えるという流れが起こり始めている。その一つは、テンソル分解を凸最適化問題として定式化するという試みである（例えば文献 [4]）。前述したようにランクを制限した行列は非凸集合であるが、トレースノルムとよばれる特異値の和に対する制約は凸集合を与えることを利用して、ランク制約をトレースノルム制約で置き換えることによって、凸最適化問題を得ることができる。同様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（三次元配列であれば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。

その他にも固有値問題を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]。凸最適化問題として定式化する方法が繰り返し固有値問題を解くことを要求するのに対して、この定式化では一般化固有値問題を1回だけ解くことによって解が得られるため非常に効率的である。

4.3 テンソル分解における外部情報の利用

上で述べたアプローチはテンソル分解の定式化を改善することでデータの疎性に起因するテンソル分解の品質低下に対処するというものであったが、多くの場合においては多次元配列として与えられる関係データ以外にも外部情報が存在する 경우가多く、これらが有用な情報として期待できる場合も多いだろう。例えば顧客と商品の購買関係において、購買情報のほか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。あるいは顧客間のソーシャルネットワークの情報も有用かもしれない。このような場合には、互いによく似た顧客・良く似た商品が似た振る舞いをするを仮定するのは妥当であろう。こういった事前知識を行列分解の最適化問題に明示的に制約として取り入れる方法 [6] や、これらの多次元配列への拡張 [7] も提案されている（図5）。結果として、行列分解や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、事前知識として類似していることが分かっているオブジェクト同士間で近くなるような解が得られるようになっている。この方法は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。

関係データが時々刻々到着するような時系列データに対して、時間情報を明示的に考慮したテンソル分解手法も提案されている。単純には多次元配列の一つの次元として時間を導入するのが一つの方法であるが、この方法は時間が連続的であるという知識を利用していないため、場合によっては利用に適さないことがある。そこで、上で述べた方法を用いて隣り合った時間同志は類似しているという制約を入れることで時間的連続性を考慮することができる。より時系列データに特化した方法としては、各時点で多次元配列が一つずつ到着するごとにテンソル分解を行うというのが提案されている[8]。これは、各時点のテンソル分解の際に、新しい解が一つ前の時間の解に近くなるような更新を行うというもので、各時点での計算コストを抑えた耐規模性の高い方法となっている。

5. お わ り に

本稿ではオブジェクト間の関係についてのデータである関係データの解析について近年の動向を紹介した。本稿では特に行列・テンソル分解を用いたアプローチに限定して扱ったが、これらのほかにも関係データのモデルとしては様々あり、すべてを網羅しているわけではないことに注意されたい。特に、行列分解の確率モデル的解釈[9]など、関係データの確率的生成モデルは本稿では取り上げなかったが非常に大切な観点であり、本稿の内容に興味を持たれた方は是非参照されることをお勧めする。関係データの解析は、古くは一階述語論理を基にした帰納論理プログラミングに遡る古典的なテーマであるともいえるが、近年のソーシャルネットワークや推薦システム等の重要性の高い応用と大規模データの出現を契機に新たな発展が急速に進んでいるテーマであり、本稿がこの活気ある領域の雰囲気少しでも伝えることができたならば幸いである。

参 考 文 献

- [1] Y. Koren, R. Bell and C. Volinsky: "Matrix factorization

techniques for recommender systems," Computer, vol.42, no.8, pp.30–37, 2009.

- [2] T. Kolda and B. Bader: "Tensor decompositions and applications," SIAM Review, vol.51, no.3, pp.455–500, 2009.
- [3] S. Rendle and L. Schmidt-Thieme: "Pairwise interaction tensor factorization for personalized tag recommendation," Proc. of the 2010 ACM International Conference on Web Search and Data Mining (WSDM), pp.81–90, 2010.
- [4] 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.
- [5] N. Nori, D. Bollegala and H. Kashima: "Multinomial relation prediction in social data: A dimension reduction approach," Proc. 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," Proc. 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, vol.25, no.2, pp.298–324, 2012.
- [8] J. Sun, D. Tao and C. Faloutsos: "Beyond streams and graphs: dynamic tensor analysis," Proc. of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.374–383, 2006.
- [9] A. Mnih and R. Salakhutdinov: "Probabilistic matrix factorization," Advances in Neural Information Processing Systems 20, pp.1257–1264, 2007.



鹿島久嗣 (Hisashi Kashima)

1997 年京都大学工学部数理工学科卒業。1999 年京都大学大学院工学研究科応用システム科学専攻修士課程修了。2007 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士 (情報学)。1999 年から 2009 年まで IBM 東京基礎研究所勤務。2009 年より東京大学大学院情報理工学系研究科数理情報学専攻准教授。2014 年より京都大学大学院情報学研究科知能情報学専攻教授。機械学習、データマイニング等のデータ解析技術の研究開発に従事。2009 年情報処理学会 長尾真記念特別賞、2012 年マイクロソフトリサーチ日本情報学研究賞、2013 年船井情報科学振興財団 船井学術賞等を受賞。