非負値行列分解　関係データ　クラスタリング

特徴的な構造を抽出するデータマイニング技術

近年，ビッグデータ解析が注目を集めています．このようなデータは人手で解析できる分量を超えているため，計算機による自動的な解析手法が必要です．本稿では，統計的機械学習に基づくデータ行列のマイニング技術を紹介します．

石黒 勝彦（いしぐろ かつひこ）／竹内 孝（たけうち こう）

NTTコミュニケーション科学基礎研究所

データマイニング技術の必要性

近年，ビッグデータを対象とした情報解析技術が大きな注目を集めています．ビッグデータのはっきりした定義はありませんが，特に注目される購買履歴データやソーシャルネットワークサービス（SNS）上のデータなどは，すでに人手で解析できる分量をはるかに超えています．

例えばオンライン通信販売サービスの巨大な購買履歴レコードはカスタマごとに管理され，商品推薦などに活かされています．一方，SNSによる簡易メッセージ（マイクロブログ）サービスであるTwitterでは，日々4.5億件以上[(1)]のツイート（メッセージ）を処理しています．

このように，人間が手動で処理できる範疇をはるかに超えたデータセットから役に立つ知識を抽出したり，あるいはそもそもデータの性質がどうなっているかを把握するためには，クラウド環境などで実装された計算機による自動計算処理，すなわちデータマイニング技術が必要となります．しかし，計算環境だけでは何もできません．問題となるのは，どのような基準で計算機に処理させるか，その設計です．

NTTコミュニケーション科学基礎研究所では，統計的・確率的な基準の意味で最適な答えを探す，統計的機械学習[(2)]に基づいたデータマイニング技術の研究開発を行っています．

多くの場合，統計的機械学習ではデータを数値化して取り扱います．本稿では，より具体的に，表計算ソフトのセルのような形，すなわちデータ行列に変換可能なデータを対象とします（**図1**）．例えばオンラインショッピングの購買履歴データは，縦軸に顧客ID，横軸に商品IDをとることで，「ある顧客が，ある商品を何度購入した」というデータ行列をつくることが可能です．また，SNS上でのユーザ間の友だち関係やフォロー関係といったリンク関係も，縦軸をリンク元のユーザ，横軸をリンク先のユーザと定義することで，「あるユーザがあるユーザを友だち・フォロー先に指定している」というデータ行列に変換できます．このように，多くのデータをデータ行列の形式に変換することができます．

本稿で紹介するのは，こうしたデータ行列に対するデータマイニング技術

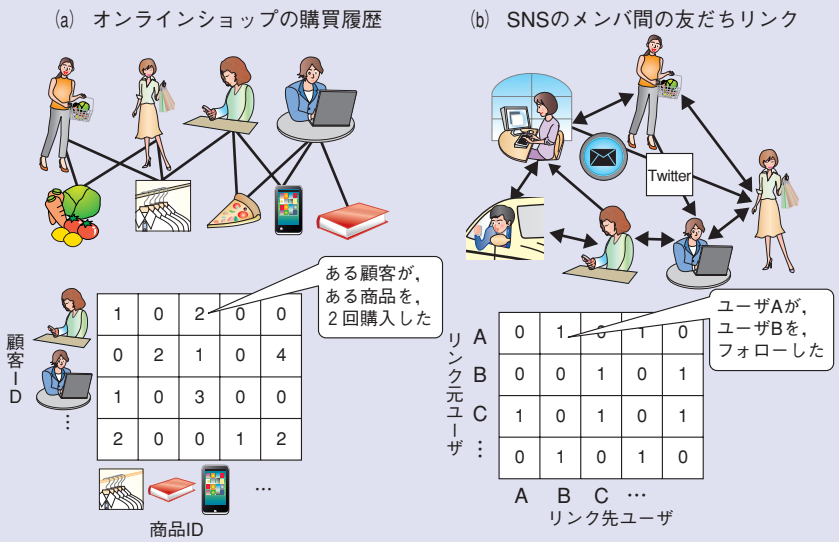

図1　データ行列形式

です．我々は統計的機械学習技術により，データ行列に潜むパターンや関係コミュニティなど，少数の特別な要素の関係として，データを構造化・説明するデータモデリング技術を研究しています．これらの手法を利用することで，事前に人手によるルール設定なしに，統計的機械学習手法により自動的に複雑なデータを少数の重要な要素に分解して，分かりやすい構造を抽出します．

非負値行列分解（NMF）による実数データ行列からのパターン発見

まず，近年多くの研究領域で注目を集めている非負値行列分解（NMF: Nonnegative Matrix Factorization）法を紹介します[3]．

NMF法は，すべての要素が非負（0以上）の実数であるデータ行列を対象とします．そして，このデータ行列を少数の「パターン」と，データごとのパターンの「強さ」を表す2つの小さい行列へ分解します．ただし，各行列の要素も非負値と制約して，元のデータ行列に対する誤差が最小になる分解を求めるのがNMF法です（**図2** (a)）．

NMFによって得られた分解の解釈は図2 (b) にあるとおりです．例えば新聞記事のデータ行列を分解することを考えると，データ行列の色分けされた各列ベクトルが各記事に対応して，各列ベクトル内の数字は，ある単語が記事内で何度使われたかを表します．このとき，NMFによって得られる分解の解釈ですが，パターンは記事の「話題」，すなわちスポーツ，経済，政治…などに対応します．パターンごとに利用される単語の頻度（強さ）は異なるので，それを見極めることで，ある新聞（緑のデータ）は，ある話題をよく報道するがほかの話題は全く出てこない，というように記事データの内容要約ができます．このパターンと，データごとのパターンの強さによる抽象化を同時に解決してくれるのがNMF法です．

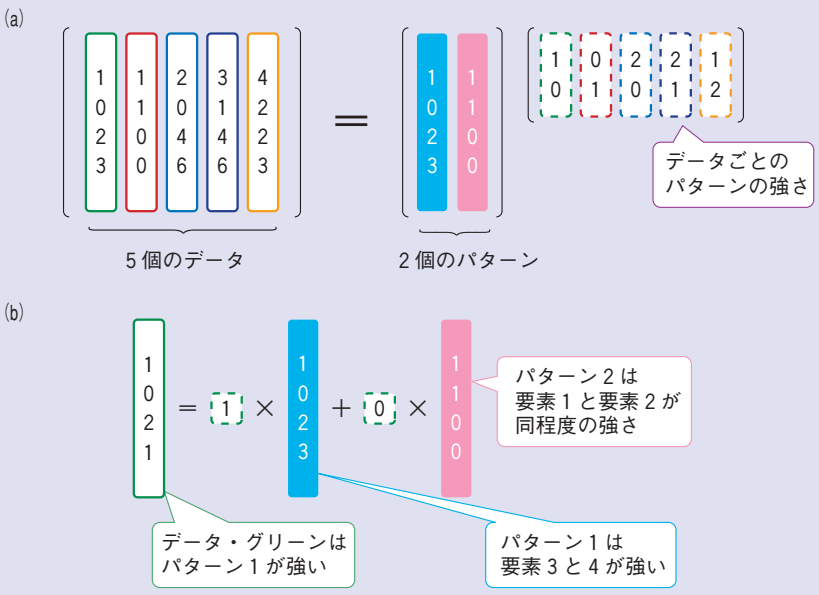

図2　NMF法の概略図

本稿では実際にTwitterからの話題抽出の例を紹介します．

Twitterではどのような話題が存在するのか，それ自体を知ることは非常に困難です．我々はTwitterのまとめサイト（一部の有志のユーザ，業者などがある特定の話題・目的で関連するツイートを集めて表示するサイト）におけるまとめ記事をNMFで解析することで，この課題を解決します．

課題では，データ行列をまとめサイトの記事（列ベクトル）と各サイト内の単語の頻度（縦の数字）で表します（**図3** (a)）．これをNMFで分解することで，話題ごとの単語のパターン（図3 (b)）と，各まとめ記事の話題の分析（パターンの強さ）を同時に計算します．

実際に2010年2月～2011年4月の間，およそ10万記事（1 000万ツイート），10万単語のデータ行列を収集して解析した話題が図3 (c) です．例えば話題1は経済の話題，話題3は宮崎での口蹄疫，話題4は東北大震災における三陸地方の津波の話題など，非常に明瞭に分かりやすい話題を発見することができました．

無限関係モデル（IRM）によるバイナリデータ行列からの関係クラスタ抽出

続いて，特に関係データと呼ばれるデータ行列に対して利用する無限関係モデル（IRM: Infinite Relational

(c)

話題	上位単語
1	経済，消費，税，政策，円，デフレ，金融，率，危機，的，増税，インフレ，財政，%，日銀，ない，所得，国債，成長，市場，銀行，兆，政府，これ，資産，投資，社会，金，言う，金利
2	映画，観る，タイトル，監督，見る，上映，級，入れる，死霊，臭い，作品，面白い，奴，シーン，版，単語，公開，映像，映画館，優勝，既存，劇場，主演，人，ストーリー，年，鑑賞，作，化
3	口，蹄，感染，処分，牛，鳥，宮崎，殺る，県，発生，農家，頭，情報，ワクチン，韓国，豚，畜産，家畜，ウイルス，日，検査，接種，インフルエンザ，市，フル，例，対策，国，確認，性
4	町，避難，情報，所，県，人，三陸，岩手，津波，市，無事，大槌，連絡，仙台，宮城，南，者，確認，被害，電話，安否，そう，状況，お願い，さん，小学校，地区，ない，来る，家
5	地震，日，県，福島，時，震度，分，発生，速報，揺れる，情報，緊急，沖，津波，震源，市，気象庁，最大，余震，人，深い，揺れ，千葉，警報，近辺，茨城，東北，宮城，被害，浦安

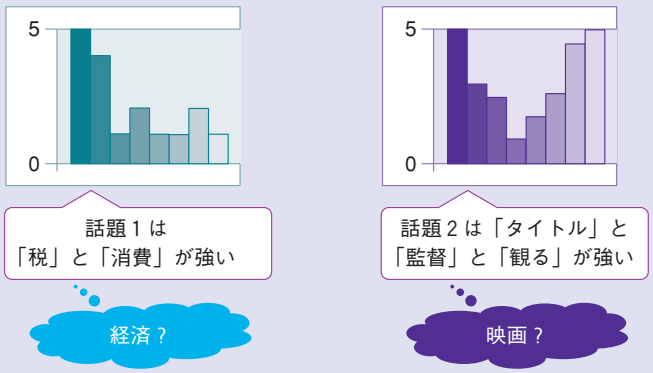

図3　NMF法のツイートデータへの適用例

Model）法を紹介します[4]．

まず，関係データを紹介します．関係データはオブジェクト間の関係の有無に着目したデータを表します．**図4**は関係データの例を示しています．関係データでは，一般に複数のオブジェクト（サーバ，顧客，アイテム等）の間のリンクの有無でオブジェクト間の関係を簡潔に表現します．

このような関係データは，**図5**に示すように要素が1（リンクあり）あるいは0（リンクなし）の二値のデータ行列に変換します．IRM法はこのようなデータ行列に対して，その行（縦方向）と列（横方向）のインデックスを入れ替えることで白黒のハッキリしたグループ分け（図中で色分けされた区画）を発見します．これは，例えば図5の例だと，「ある特定の顧客グルー

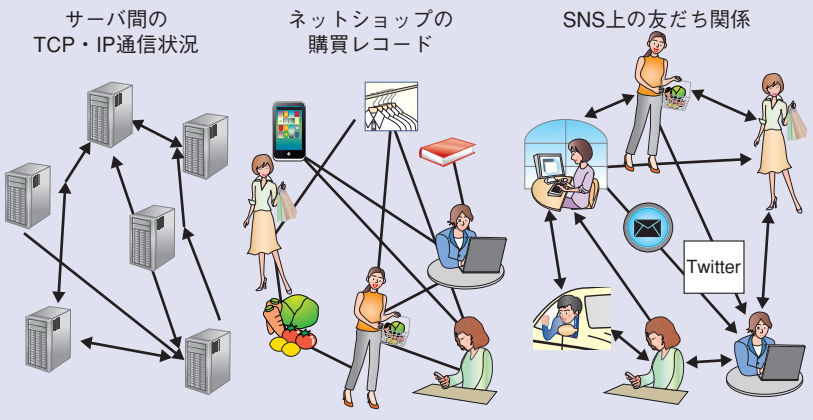

図4　関係データの例

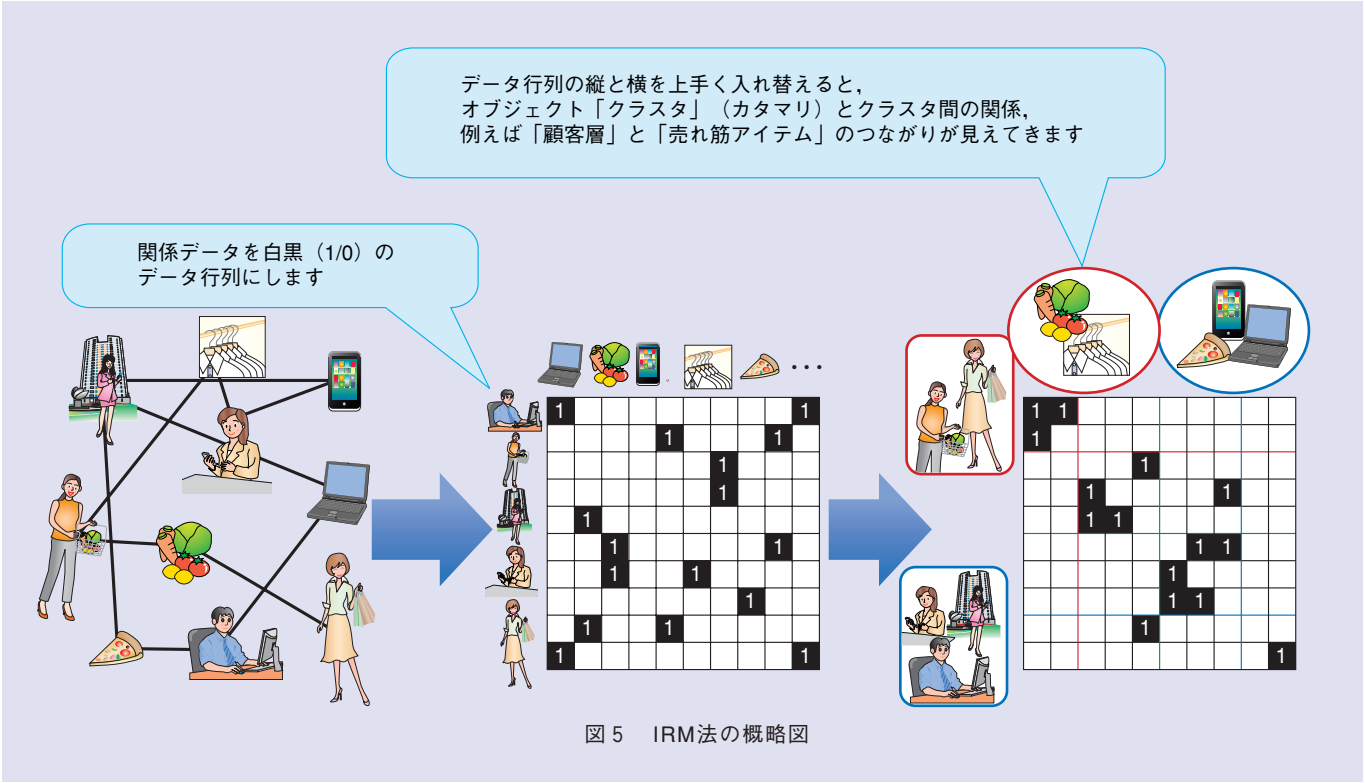

図5　IRM法の概略図

プ」に対して「よくお買い上げいただいている商品群」という，「グループ×グループ」の購買パターン構造を発見するといった用途に利用できます．あるいは，例えばSNSの友だち関係データに適用すると，SNS内の閉じた関係，すなわちコミュニティの発見とコミュニティ間のインタラクションの解析が可能となります．

IRM法の1つの特徴は，グループ分けのパターン数を統計的に最適であるか，という観点にしたがって自動的に決定してくれる点にあります．したがって，ユーザは1あるいは0の要素を持つ関係データ行列を入力すれば，自動的に以上のような解析を行うことが可能となります．

このIRM法はNTTコミュニケーション科学基礎研究所と海外の大学の共同研究で開発された手法ですが，ここ数年でさらに我々のグループ特有の機能拡張が実現されています．ここではIRM法を拡張した2種類の方法を紹介します．

⑴　時間変化関係解析法

私たちの独自技術[5]では，時間変化する関係データに対しても上記の処理を行うことができます．

一般のいろいろな関係は，時間とともに変化するのが普通です．例えば，ある組織内での人間関係は日常徐々に変化するとともに，組織変更などで急激に変化します．我々はこのような関係の時間変化を表現する機械学習モデルを提案しました．実験では，エンロン社の社内Eメールの履歴を解析することで，例えば金融・財務関係の社員のコミュニティを発見・追跡したり，あるいは瞬間的に社内全体に大きな影響を与えた時の人を検出するなど，エンロン社内の人間関係の時間変化を容易に追跡できるようになりました．

⑵　スパースネス対処法

最新の成果[6]では，実際の購買履歴データなどの解析で派生するスパースネスの問題に対処する拡張手法を発表しています．

ネットにおける購買履歴データの多くは，非常にたくさんの「ユーザ×商品」の組み合わせに対して，実際に購買行動が観測される組み合わせはごく少数です．このようなデータをデータ行列にすると，ほとんどの要素が0となるスパースな行列になります．このようなデータ行列に単純にIRM法を適用すると，**図6**⑷のように，たくさんの関係のブロックが抽出されます．こ

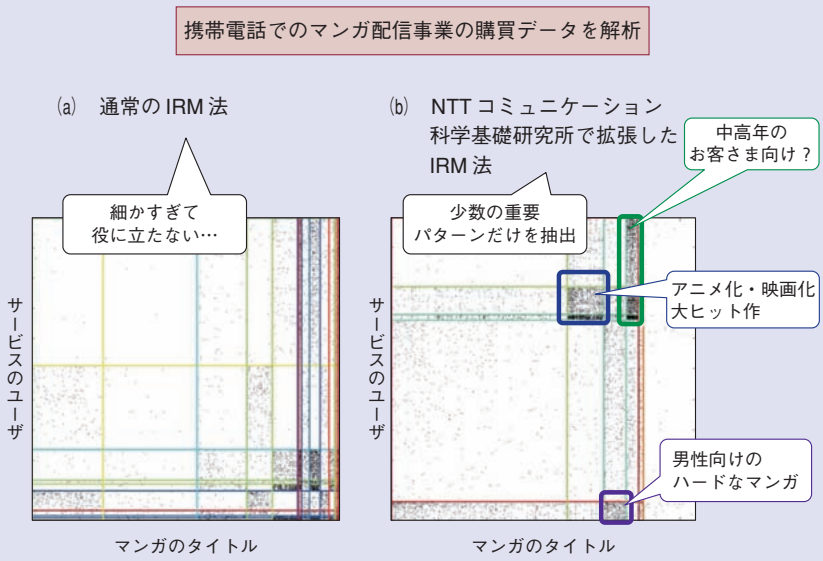

図6　スパースネス対処を施したIRM法の適用例

のままでは，購買行動をパターン化してもそのパターンを実際に人間が解釈する際には多くのコストがかかってしまいます．

そこで，我々は重要な購買パターンだけをIRM法によって抽出する方法を考案しました．アイデアとしては，「皆が買っている，あるいは誰も買っていない商品はマイニングしても価値がない」という仮説です．これは逆もまた然りで，「何も買わない，あるいは何でも買うユーザ」もマイニングの目的としてはあまり価値がありません．このように，特定のパターンを持たないオブジェクトをデータ行列から除外し，残りのオブジェクトだけでIRM法を適用すると，図6(b)のように，少数の，かつ直感的に人間が理解できる購買パターンを発見することができるようになりました．

今後の展開

本稿では，人手による解析が困難な大量・複雑なデータ解析に有用なデータマイニング技術である，NMF法とIRM法を紹介しました．

このような手法に限らず，さらに複雑な関係を表す高階なデータやストリームデータのモデル化，高速探索，予測手法など，統計的機械学習手法の守備範囲は日進月歩で拡大しています．NTTコミュニケーション科学基礎研究所では先端の基礎技術の研究にコミットし続けることで，グループ全体のデータマイニング技術の見通しを広げ続けていきます．

■参考文献

(1) https://twitter.com/
(2) ビショップ：“パターン認識と機械学習―ベイズ理論による統計的予測，”シュプリンガー・ジャパン，2007.
(3) M.Mørup: “Applications of Tensor (multiway array) Factorizations and Decompositions in Data Mining,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1, No. 1, pp. 24-40, 2011.
(4) C.Kemp, J.B.Tenenbaum, T.Griffiths, T.Yamada, and N.Ueda: “Learning Systems of Concepts with an Infinite Relational Model,” AAAI-06, Boston, U.S.A., July 2006.
(5) K.Ishiguro, T.Iwata, N.Ueda, and J.Tenenbaum: “Dynamic Infinite Relational Model for Time-varying Relational Data Analysis,” NIPS 2010, Vancouver, Canada, Dec. 2010.
(6) K.Ishiguro, N.Ueda, and H.Sawada: “Subset Infinite Relational Models,” AISTATS 2012, La Palma, Spain, April 2012.

（左から）竹内　孝/ 石黒　勝彦

データ行列形式に変換できないデータであっても，それらに対応する機械学習技術は多数開発されています．我々のグループでは，本稿では書ききれないほどさまざまな形式・問題・対象に対する機械学習技術を研究開発しています．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
協創情報研究部
知能創発環境研究グループ
TEL　0774-93-5399
FAX　0774-93-5155
E-mail　ishiguro.katsuhiko@lab.ntt.co.jp