ビッグデータを活かす機械学習技術

うえだ　なおのり
上田　修功
NTTコミュニケーション科学基礎研究所 所長※

現在，想像を超える勢いでデジタルデータが生成，蓄積されています．まさにビッグデータ時代の到来です．そのような膨大かつ多様なビッグデータから新価値を生み出すための分析技術として，機械学習技術が有望視されています．本稿では，機械学習技術とはどのような技術か，また，なぜビッグデータ分析に機械学習技術が有用かについて概説し，さらにビッグデータ分析に対するNTT R&Dでの今後の取り組みについて紹介します．

ビッグデータ時代の到来

ビッグデータとは，既存のデータベースでは収集・編集・蓄積が困難な，膨大かつ多様な非構造化データと定義されていますが，さらに，これらのデータから新価値を生み出し，産業・社会に大きな影響を及ぼし得るという意味も込められています．実際，ビッグデータというバズワードの生みの親といわれるマッキンゼーの調査部門は，ビッグデータが保健医療，小売り，製造業などにおいて莫大な金額の価値創出につながる可能性を示唆しています(1)．

ビッグデータの経済価値を受けて，米国オバマ政権は，2012年３月29日にビッグデータ関連の研究開発に総額２億ドル以上を投じるという研究開発イニシアティブを発表しました．具体的には，米国国立科学財団（NSF），米国国立衛生研究所（NIH），米国大統領府科学技術政策局（OSTP），米国国防総省（DoD），米国エネルギー省（DoE），米国地質調査所（USGS）の６つの政府機関が連携主導して，ビッグデータの管理・分析のための最先端中核技術の発展の促進，科学技術分野での発見の加速，国家安全保障の強化，さらにはビッグデータ技術分野の発展・活用に必要な人材育成を行うために２億ドル以上の投資を行うとのことです．つまりビッグデータはビジネスの世界だけではなく，研究面でも重要分野として位置付けられています．

ビッグデータ分析の特徴

既存のビジネスインテリジェンスでは，主として，あらかじめ分析シナリオが決められた社内データや顧客データなどの構造化データを対象としていましたが，ビッグデータでは，センサデータ，ソーシャルメディアデータ，機械のログデータなど多様な非構造化データが対象となります．さらに，蓄積型のデータだけでなく，逐次的に生成されるストリームデータも分析対象となります．

また，分析方法も多様です．例えば，インディアナ大学のボーレン准教授が約270万人のTwitterユーザの約980万回のツイートを分析し，株価を約87％の精度で予測し，関係者を驚かせました．精度もさることながら，従来の金融工学的なアプローチではなく，テキスト自動分類という機械学習技術による分析という，当該分野では全く斬新な分析シナリオだったことも興味深い点です．

ビッグデータ分析では，使用するデータもその分析のために作成されたデータソースとは限らず，どのようなデータでどのように分析するかという分析シナリオのデザインが重要です．ビッグデータ分析における新価値創造は，従来の「仮説検証型」の分析では不十分で，仮説そのものを多種多様な情報から発見する「仮説発見型」の分析により，初めて実現可能になります．

※　現，NTTコミュニケーション科学基礎研究所 機械学習・データ科学センタ長

機械学習技術とは

機械学習技術とは，一言でいうと，情報システムに学習能力を持たせる技術です（**図1**）．学習能力においては，学習時の所与のデータだけでなく，未知のデータでも性能を発揮する汎化能力が重要です．試験勉強で例えると，いくら参考書で勉強しても本番のテストで点が取れなければ学習したとはいえないのと同じです．

機械学習研究は，1950年代に人間の知能獲得の模倣を目標として開花した人工知能研究がそのルーツといわれています．そして現在の機械学習技術は，2000年ごろから研究が盛んになった統計的機械学習技術が主流で，人間の知能獲得の模倣ではなく，数理統計・最適化理論を土台とする工学的アプローチへと変遷しています．その意味では機械学習は古くて新しい研究分野といえます．

機械学習の主な枠組みとして，①教師あり学習，②教師なし学習，③半教師あり学習，④アンサンブル学習，が挙げられます．①は，例えばパターン認識応用では，観測データに対し何らかの特徴抽出を行って得られた特徴ベクトルと，そのデータのクラス（グループ）ラベルが与えられ，学習によって特徴ベクトルと，クラスラベルとの関係を学習し，クラスラベルが未知のデータに対して学習済の学習器でそのクラスラベルを予測するというものです．

一方②は，特徴ベクトルのみから，類似したクラスを学習します．これはクラスタリングとして広く知られている学習技術です．

①の教師あり学習の場合，クラスラベルデータの収集が実用上の問題となります．例えば，Webテキストからの評判分析タスク（意見がポジティブか，ネガティブか，いずれでもないか）の場合，あらかじめ人がテキストを読んでクラスラベルを付与しなければならないため，労力と時間がかかるという実用上の問題が生じます．この課題解決のための学習法が③の半教師あり学習です．③では少数のラベルありデータに大量のラベルなしデータを混合してクラス分類性能を向上させる学習法です．直観的には，**図2**に示すように，少数のラベルありデータだけではクラス境界が曖昧であっても，大量のラベルなしデータをラベルありデータと混合することで，あたかもラベルが付与されているかのごとく，データのクラスタリング効果でよりクラス分類境界が明確化されるわけです．

④のアンサンブル学習とは，人間で例えるならば，分からないときは皆で議論するという学習法です．個々に学習した学習機械を複数用意して，未知データに対しては，複数の学習機械の出力結果をまとめて（例えば単純な多数決），最終的なクラスを決定する学習法です．一般に，単一の学習機械での性能向上には限界があります．ア

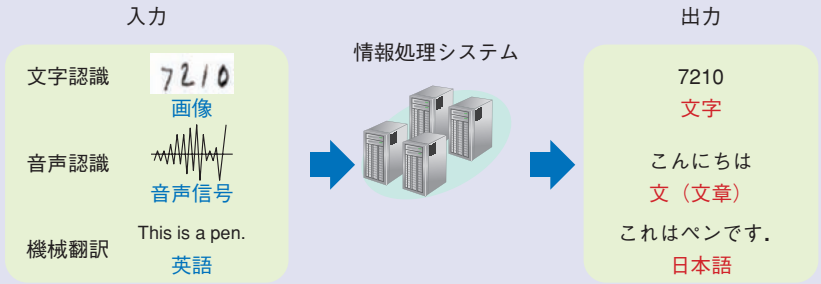

図1　機械学習技術

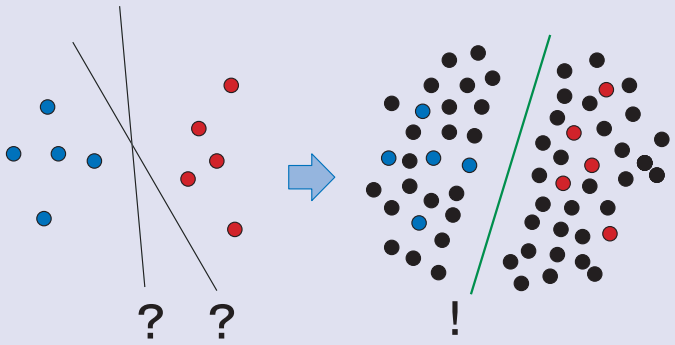

図2　半教師あり学習

ンサンブル学習は汎化性能向上のための実用手法といえます．アンサンブル学習は，クラウドソーシングなどの集合知からのデータ分析にも今後の発展が期待されています(2)．

上記学習の枠組みはパターン認識の例で説明しましたが，応用によっては，シンボリックなクラスラベルだけではなく，数値データを教師データとすることもあります．

なぜ機械学習が有用か

前述したように，機械学習技術は特定の応用分野に依存しない汎用的な技術で，観測データと目標タスクさえ明確化すれば，適切な学習スキームに応じて分析結果を得ることができます．この汎用性・柔軟性が多種多様なデータ分析を必要とするビッグデータ分析に適している理由の１つです．

また，ビッグデータ分析ではノイズの取り扱いも重要です．ノイズとはセンサなどの観測ノイズだけではなく，ソーシャルデータなどでの内容の信憑性・信頼性もノイズの一種です．このようなデータに対してはその信頼性を確率的に評価可能な技術が重要で，統計的機械学習技術はまさにそれらに適した技術です．統計的機械学習アプローチでは，**図３**に示すように，観測データは何らかの確率モデルから生成されていると仮定します．ここでの確率モデルは必ずしも真理の解明のためのモデル化である必要はなく，観測データをその生成モデルに従って生成できれば良いという工学的な考え方に基づきます．

確率モデルは，一般にはパラメータで特定される確率分布に相当しますが，重要な点は本来観測されない潜在変数の導入がモデルの自由度を高め，より精度の高い分析を可能にすることです．学習は生成の逆問題で，仮定したモデルのパラメータ，潜在変数の値を観測データから推定する問題に相当します．学習（推定）過程では，ベイズの定理*1に基づいて，観測データが得られたもとでのモデルパラメータと潜在変数の事後確率分布を計算し，その事後分布を用いて，将来のデータを予測します．近年では，潜在変数の個数もデータから自動決定可能な理論的研究も，統計的機械学習分野で精力的に研究されています(3)．

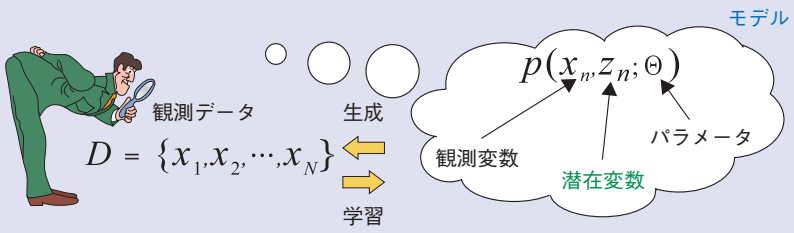

図３　統計的機械学習におけるデータ生成と学習のプロセス

潜在変数の直観的理解のために，レコメンデーションシステムの中核技術である協調フィルタリング（CF: Collaborative Filtering）技術を用いて説明します．CF技術とは，蓄積された多くの人の嗜好情報を土台に，ある人の嗜好を，その人の嗜好と類似した他人の嗜好情報から自動的に推測する技術です．あるターゲット顧客に対し，あるターゲット商品についての評点を推測する場合，ほかの顧客の中からターゲット商品以外の商品でターゲット顧客と評点が類似している顧客を探し，それら類似顧客のターゲット商品に対する評点に基づいてターゲット顧客の評点を予測する，という古典的な手法でした．

一方，潜在意味構造解析と呼ばれる統計的機械学習アプローチでは，まず，顧客にはコミュニティが存在すると仮定します．このコミュニティは観測されないので潜在変数に相当します．そして，顧客と商品とをコミュニティを介して関連付けることで顧客と商品間の膨大な組み合わせの問題を緩和することができ，評点予測精度を格段に向上することが可能となっています．

このように，観測データの生成モデルにおいて，観測変数だけでなく，潜在変数を導入することでより自然かつ柔軟なモデル化が実現できます．換言すれば，潜在変数の導入の仕方がビッグデータ分析シナリオそのものであり，分析シナリオにおいてその学術的な信頼性の根拠を与える核技術が統計的機械学習技術といえるのです．

関係データ解析

ビッグデータ解析では，ヒューマンネットワークや購買ログデータのように，グラフ，もしくはその等価表現である行列で表現されたデータからのクラスタリング分析が核技術として有用

＊1　ベイズの定理：条件付き確率に関し，トーマス・ベイズによって示された定理．

です（**図 4**）．これまでNTTコミュニケーション科学基礎研究所では，このような関係データからのクラスタリング技術をいち早く手掛け，実データでの有効性も検証してきました[(4)]．

次に，携帯電話でのコミック配信サービスの購買データの分析結果を紹介します（**図 5**）．オリジナルデータは，図 4 の購買ログデータと同じ形式で，顧客があるコミックを購買したか否かで，1 または 0 の値が行列の要素として表現されています．この行列の行と列を適切に並び替えることで，図 5 に示すように，どのような顧客がどのようなコミックを主に購買したかが一目瞭然となります．

次に，行列の並び替えを機械学習技術でどのように実現しているのかについて概説します．簡単にするため，図 4 のヒューマンネットワークの行列データのような行と列がともに同一オブジェクトの場合（この例の場合は行も列もユーザ）について説明します．生成モデルアプローチでは，前述したように，観測データがどのようなプロセスでデータが生成されたかをモデル化します．例えば**図 6** (d) の観測データでは一見関係性が曖昧ですが，これは仮定した生成モデルで得られたクラスタリング結果（図 6 (c)）をインデッ

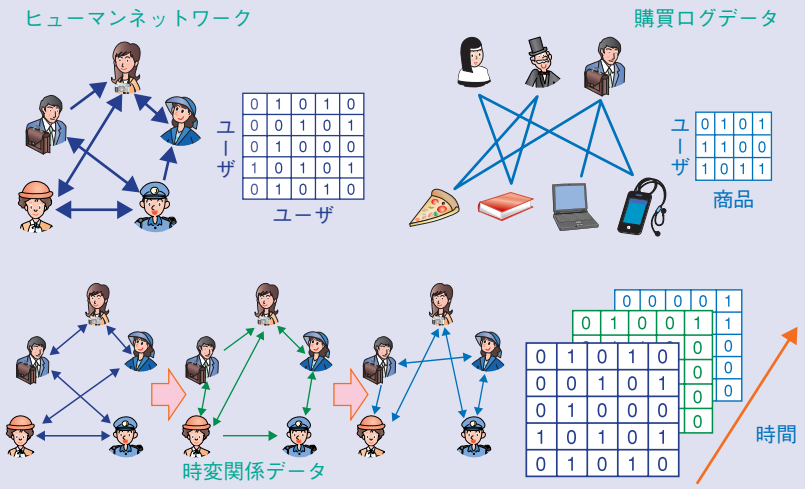

図 4　さまざまな関係データとその行列（テンソル）表現

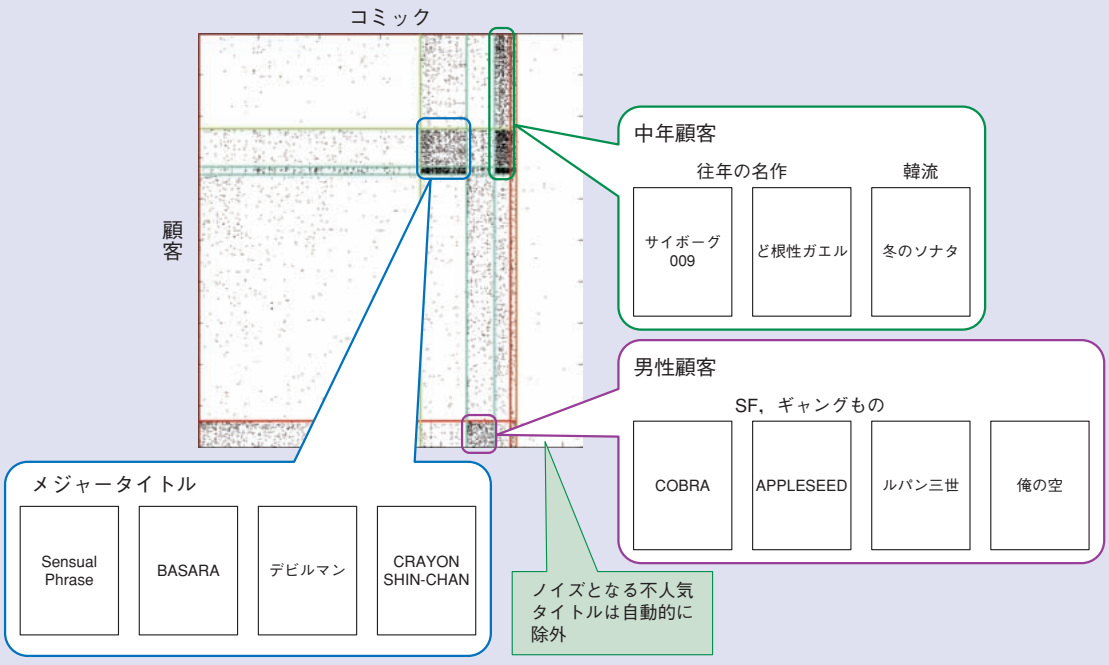

図 5　顧客とコミックの同時クラスタリングの結果

クスで並び替えたものと解釈します．

では図 6 (c) がどのような手順で得られたかというと，まず，分割の事前分布であるディリクレ過程*2という確率モデルを用いて 9 個のオブジェクトを分割します(3)．その結果，図 6 (a) のように，3 つのクラスに分割されたとして，次に，各クラスの組で関係性の強さをモデル化します．例えば，図 6 (b) の行でのクラス 1 に属すオブジェクトが列でのクラス 2 に属すオブジェクトに対し，0.9の確率で 1 を生成（等価的に0.1の確率で 0 を生成）するというふうにクラス間での関係を表す確率値を定めます．そしてその確率に従って，1 または 0 を生成すると図 6 (c) が得られるわけです．

当然ながら，実際には図 6 (d) の観測データしか得られないわけで，分割結果や確率値は未知ですので，学習過程では図 6 (d) からベイズ推定に基づいて図 6 (a) のクラス分割を推定するという逆問題を解くことになります．具体的には，図 6 (d) が与えられたもとでオブジェクト分割の事後確率分布を最大化するオブジェクト分割を求める問題となります．この逆問題は解析的には解くことはできず，マルコフ連鎖モンテカルロ（MCMC: Markov Chain Monte Carlo methods）法や変分ベイズ法を援用して近似解を求めることになります．本アプローチでは，あらかじめクラス数を決めておく必要がない点，並び替えのすべての組み合わせを考慮する必要がない点で，従来のクラスタリングアルゴリズムと本質的に異なります．より詳しい内容については文献 (3)～(5) をご参照ください．

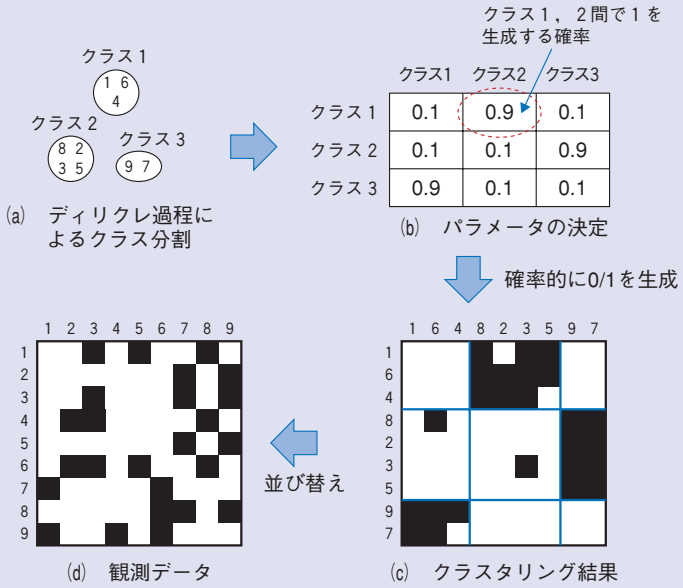

図 6　関係データの生成モデルの考え方

今後の展開

本稿では，ビッグデータ分析の核技術である機械学習技術，特に統計的機械学習技術に焦点をあて，その考え方と分析例の一端を紹介しました．ビッグデータ分析に，機械学習技術だけでは不十分で，NTT研究所の関連技術（データベース技術，並列分散計算技術），さらにセキュリティ，トラフィック，CRMなどの応用領域の研究者と密に連携し，ビッグデータの研究開発に取り組んでいきます．

*2　ディリクレ過程：個のグループ化に対する事前分布．グループ内のメンバ数が多いグループほど，確率的にさらにメンバ数が増大する確率過程．

■参考文献

(1) MGI report：“Big data: The next frontier for innovation，competition，and productivity，” 2011.
(2) スロウィッキー：“「みんなの意見」は案外正しい，” 角川文庫，2009.
(3) 上田・山田：“ノンパラメトリックベイズモデル，” 応用数理，Vol.17，No.3，pp.196-214，2007.
(4) 石黒・竹内: “特徴的な構造を抽出するデータマイニング技術，” NTT技術ジャーナル，Vol.24，No.9，pp.14-18，2012.
(5) K.Ishiguro，N.Ueda，and H.Sawada：“Subset Infinite Relational Models，” AISTATS 2012，La Palma，Canary Islands，April 2012.

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
企画担当
TEL　0774-93-5020
FAX　0774-93-5015
E-mail　cs-liaison@lab.ntt.co.jp