ビッグデータ時代の最先端データマイニング

藤巻 遼平・森永 聡

要　旨

近年のビジネスでは、ビッグデータを分析して得られる知見を活用することが不可欠になりつつあります。しかしビッグデータは、異なるパターンや規則性に従っているデータが混在して収集・蓄積されていること（データの異種混合性）が多いため分析が難しく、そのため、データの異種混合性が非常に重要視されるようになりました。本稿では、NECが開発した異種混合データの最先端分析技術である「異種混合学習」について、いくつかの応用事例を交えながら紹介し、これまで漫然と収集されているだけだったデータ活用の可能性に関して論じます。

キーワード

●ビッグデータ　●異種混合データ　●データマイニング　●機械学習
●異種混合学習　●因子化漸近ベイズ推論

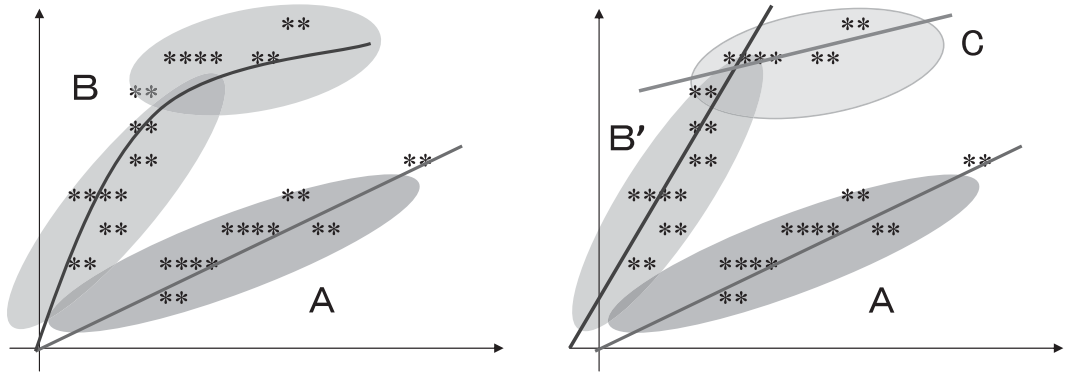
1. まえがき

近年のビジネスでは、ビッグデータを分析することにより、頻出するパターンや隠れた規則性などに関する知識を自動的に発見し、価値ある情報として有効活用することが不可欠になりつつあります[*1]（このような知識発見の技術を一般的にデータマイニングと呼びます）。例えば、電力需要の予測のために、気温などの各種センサデータの値と電力需要量の動きの間に成立する規則性を抽出したうえで、随時、その規則を現状のセンサデータに適用し、その後の需要量を導出するといったことなどが行われています。

ビッグデータ分析の難しさとして、その規模が大きいこととともに、収集や蓄積されているものの中に、異なるパターンや規則性に従っているはずのデータが混在してしまっている（データの異種混合性）という問題が挙げられます（異種混合データの問題）。特に、複雑な異種混合データでは、単にデータに複数のパターンや規則性が存在するだけではなく、各パターンの性質が大きく異なるという特徴があります（**図1** 左は線形な性質と非線形な性質が混在する異種混合データの模式図）。 電力需要を例にすると、何らかの要因でセンサ値と電力需要の関係性が切り替わっているのに、従来の分析技術ではその様子をうまく捉えられずに、予測精度が低下するという問題がありました。対処方法としては、曜日や時間帯といった、規則性が切り替わる要因を専門知識などに基づいて試行錯誤で想定し、その単位にデータを分割して別々にそれぞれ単一規則を自動抽出するということが多く行われていました。しかし、専門家でも適切に要因想定を行うことは非常に難しく、データ分割が不十分で異種混合性が解消できない、あるいは、データ分割が過剰でパターンが断片化する（どちらも予測精度の低下の主因となる）、といった問題が生じていました。

本稿では、まず異種混合データ分析の難しさについて説明します。一言で述べれば、データ分割候補の可能性数が膨大であるため網羅的な探索をすることが現実的にはできない点が、本質的な難しさを象徴しています。次に、NECが開発した異種混合データの最先端分析技術である異種混合学習について紹介します。この技術は、「因子化漸近ベイズ推論」という高度な機械学習技術を応用したもので、本稿ではその基本的な考え方を紹介します。最後に、異種混合学習の応用例として、ビルの電力需要予測の実証実験を紹介します。異種混合学習技術によって、従来の異種混合データを想定しない予測方法に比べて7.6ポイント（10.3%→2.7%）、専門家によるデータ分割に依存する方法に比べて2.1ポイント（4.8%→2.7%）、予測精度が改善しました。

[*1] 矢野経済研究所の2012年調査[1]では、2011年度のビッグデータ市場規模は1,900億円、2020年度には1兆円を超えると予測しています。



線形モード(A) ＋ 非線形モード(B)　　線形モード×3(A, B', C)

図1　異種混合データの模式図。データを適切に分析するためには、膨大なデータ分割の候補集合から、適切な分割方法を発見する必要がある（楕円がデータ分割方法、線が予測モデルに対応する）

2. 異種混合データ分析の難しさ

異種混合データを精度良く分析するために最も重要な点は、同一のパターンや規則性を持つグループごとにデータを分割し、内在する異種混合性を解消することです。しかし、データ分割の仕方には膨大な可能性（場合によっては無限の候補）があり、これをしらみつぶしに検証していくことは現実的には不可能です。データを複数のグループに分割する際の論点には大きく以下の3つがあります。

1) グループの数（どのくらい混ざっているのか？）
2) 各グループの分け方（どのように分離するか？）
3) 各グループの性質に応じた適切な予測モデル

これらの論点は、独立して、あるいは1）から3）へ順番に決定できるものではなく、相互の依存関係を考慮して同時に決定する必要があります。例えば、データに非線形な関係性と線形な関係性が混在しているという仮説に対しては（図1左）、データを2つのグループ（Bの楕円とAの楕円）へ分けることで精度の良い予測モデルを得ることができますが、複数の線形な関係性が混合しているという仮説に対しては（図1右）、グループ数は3が適切となります。また、図1の左右どちらの場合においても、各グループの分け方（楕円）は、（非）線形な関係性（予測モデル）が適合するデータの集まりによって決まるため、1）や3）を無視して2）を決めることはできないことが分かります。

では、1）から3）の論点を同時に考えた場合の、データ分割の候補数は、具体的にどのくらいの数があるのでしょうか。例として、センサと電力需要の値を大量に蓄積したビッグデータを分析して、それらの間の隠れた規則性を見つけるケースを考えます。更に、問題の本質を説明するために、ここでは予測モデル（電力需要の予測式）の候補は、説明変数（センサ値）の二次式で表せる場合に限定するとします。説明変数の数（センサ数）が10個、予測モデルに利用可能なセンサ数を3個、データ分割後のグループ数を4に固定した場合には、概算で$({}_{100}C_3)^4 = 6.84 \times 10^{20}$（10の20乗は、1兆の1億倍）の予測モデル候補が存在します。より複雑なケースでは、ほとんど無限に近い数のデータ分割と予測モデルの組み合わせ候補があり、単純なアルゴリズムでは探索に非現実的な時間が掛かってしまうことが分かります。

第1章で述べたように、従来はこの問題を解決するために、規則性が切り替わる要因を専門知識などに基づいて試行錯誤で想定し、その単位にデータを分割して、グループごとに単一規則を自動抽出するということが多く行われていました。しかし、複雑なシステムから取得されるデータに対して適切なデータ分割方法を発見することは、専門家にも簡単なことではなく、不適切な分割によって予測精度が低下する、適切な分割方法を見つけるための試行錯誤に膨大な工数が必要となる、といった問題がありました。

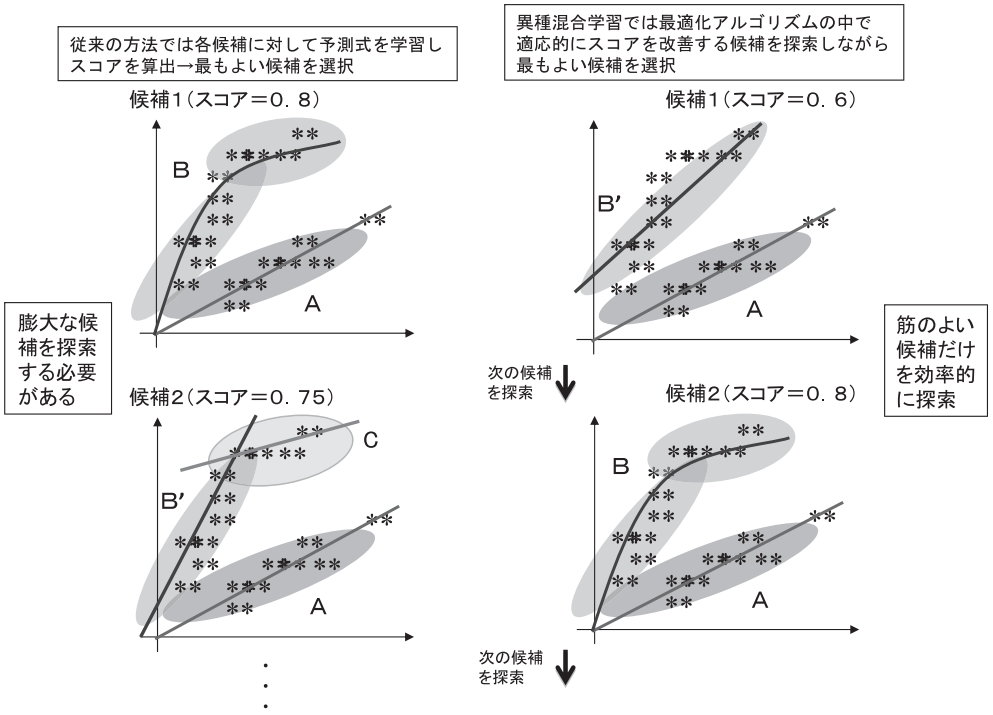


図2　異種混合学習と従来法によるデータ分割及び予測モデルの探索方法の違い

3. 異種混合学習によるデータマイニング

弊社では異種混合型データをマイニングするために、新たに異種混合学習技術を開発しました。この技術は、第2章で述べた1）から3）の3つの論点を、データ分割や予測モデルの組み合わせ爆発の問題を回避して、高速に最適化することが可能です。

図2 を用いて、従来の方法（例えば交差検定法やベイズ情報量基準など）による学習と異種混合学習の違いを説明します。従来の方法では、各モデル候補に対してスコア（情報量基準）を算出し、最もスコアの良いモデルを選択します。しかし、第2章で述べたように、異種混合型データの学習ではモデル候補数が膨大なため、非現実的な計算時間が掛かってしまいます。一方で、異種混合学習では、分割数、分割方法、グループごとの予測モデル1）から3）の論点を同時に適応的に探索することで、無駄な候補を探索することなく筋の良い（予測精度が高い）モデルだけを調べ、最適なデータ分割と予測モデルの発見をすることができます。

異種混合学習の高度な探索・最適化の背後には、「因子化漸近ベイズ推論」[2) 3) 4)] という最新の機械学習理論があり、以下のような性質が異種混合学習の根幹を支えています。

(1) 因子化情報量基準：

機械学習分野において複数のモードを持ったモデルは「非正則（特異）モデル」と呼ばれ、従来の情報量基準（ベイズ情報量基準や赤池情報量基準など）では適切にデータ分割や予測モデルの良さを測ることができないことが知られています。異種混合学習では、因子化情報量基準という非正則モデルに対する独自の基準によって適切なデータ分割や予測モデルを選択します。

(2) 適応的探索アルゴリズム：

図2で説明しているように、異種混合学習では、グループ数、分割方法、グループごとの予測モデルを適応的に

変更しながら探索します。この探索の際に、特殊な探索手法を用いることで、変更後のモデルが因子化情報量基準の意味で、必ず前のモデルよりも良くなることを保証しています。前のモデルより必ず良くなるモデルを適応的に選択することができるため、それよりも良くないモデルを探索する必要がなく、膨大な候補から高速にデータ分割と予測モデルを発見することができるというわけです。

(3) 調整パラメータ（属人的要素）の排除：

多くの機械学習・データマイニングアルゴリズムには、分析者が手動で調整すべきパラメータが存在します。しかし、この調整はアルゴリズムに対する数理的理解が必要となり、一般的には非常に高度なスキルを必要とします。異種混合学習では、従来必要だった調整パラメータを因子化漸近ベイズ理論よって決定します。これによって属人的要素を排除し、分析を自動化しています。

(4) モデルの同定性：

データ分割と予測モデルの候補中には、図1の左右に示されるように、性能が極めて近い（あるいは完全に等価な）モデルが存在します。このような等価なモデルが存在する場合には、モデルをうまく学習することができないという問題（モデルの非同定性の問題）が知られています[5]。異種混合学習は、このように等価なモデルが存在する状況において、モデルを一意に特定する「モデルの同定性」を有していることが理論的に示されています。

4. 電力予測に関する実証実験

異種混合モデルの効果を確認するために、ビルの電力需要予測を例に実証実験を行いました。近年の世界的な燃料価格の高騰を背景に、電力需要を正確に予測し、単純なピークカットのみではなく、よりインテリジェントな制御をするこ

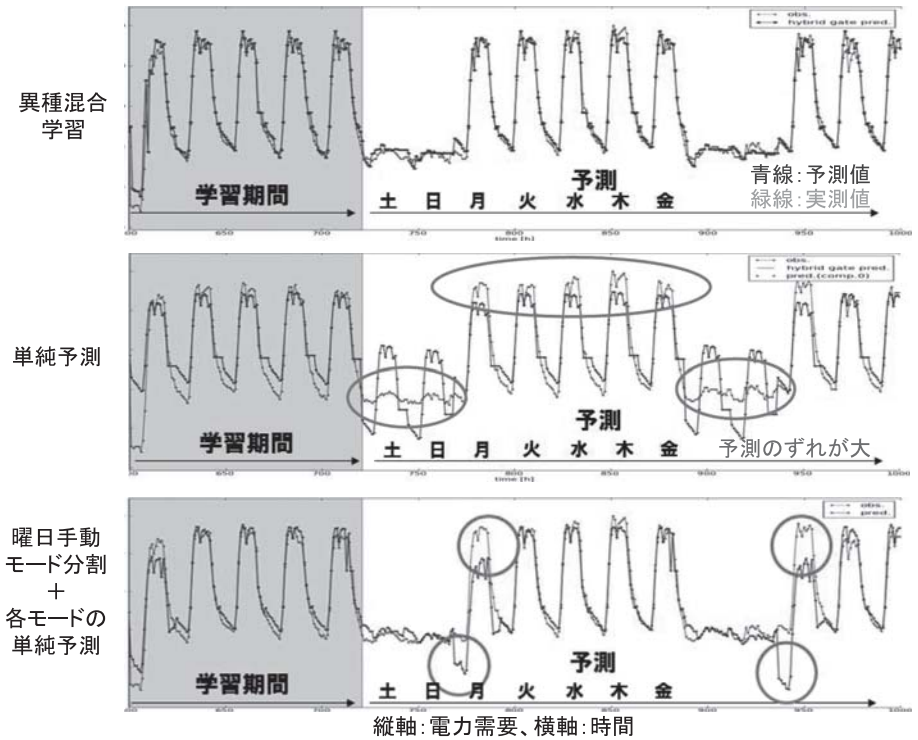


図3　ビルの電力需要予測に対する異種混合学習の適用結果

とで、エネルギーコスト削減に貢献することが可能と考えられます。

図3 は、異種混合学習による予測（上段）、単純予測（従来の機械学習手法による予測：中段）、手動データ分割予測（曜日ごとに電力需要傾向が異なると仮説を立て、曜日ごとにモデルを学習させた予測：下段）をそれぞれ表します。各手法の予測誤差（線のずれ）を観察すると、中段や下段では需要の傾向が切り替わる点をうまく学習できず、予測のずれが大きくなる期間（丸囲み）が存在します。一方で、異種混合学習では、複数のモデルを自動的に切り替え（この例ではデータが3つに分割されています）、傾向の切り替え点においても精度良く予測ができていることが確認できます。異種混合学習と単純な予測では7.6ポイント（10.3%→2.7%）、手動モードに比べて2.1ポイント（4.8%→2.7%）、予測精度が改善しました。

5. 広い適用可能性

本技術を、例えばビルの電力需要の予測に活用した場合、外気温・曜日・時間帯などと電力消費量の関係が一定していないビルにおいても、収集したデータに混在するいろいろな規則性を発見して活用することで、高精度な予測を行うことができます。また、医療領域に活用した場合、日常生活において収集しているデータから異常パターンを発見することで、見つけるのが難しい病気の早期発見に貢献することが期待できます。

6. おわりに

本稿では、ビッグデータ分析の最先端技術である異種混合学習技術について、異種混合型データ分析の難しさ、異種混合学習の基本的な考え方や性質、電力需要予測に関する実証実験結果を紹介しました。

ビッグデータ分析の重要性の高まりとともに、異種混合データのマイニング技術は今後ますます重要になり、異種混合学習の応用範囲は更に広がっていくと考えられます。

参考文献

1) 矢野経済研究所:“2012 ビッグデータ市場ー将来性と参入企業の戦略ー,” C54101300, 2012.4
2) R.Fujimaki , S.Morinaga:“Factorized Asymptotic Bayesian Infer - ence for Mixture Modeling,” JMLR W&CP 22: 400-408, 2012
3) R. Fujimaki,K. Hayashi:“Factorized Asymptotic Hidden Markov Models,” Proceedings of the 29th International Conference on Ma - chine Learning (ICML), 2012
4) R. Fujimaki,Y. Sogawa,S. Morinaga:“Online heterogeneous mixture modeling with marginal and copula selection,” Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp.645-653, 2011
5) S. Watanabe:“Algebraic geometry and statistical learning” Cam - bridge University Press, 2009

執筆者プロフィール

藤巻 遼平
NECラボラトリーズアメリカ
メディアアナリティクス部門
リサーチスタッフメンバー
博士（工学）

森永 聡
中央研究所
情報・ナレッジ研究所
主任研究員
博士（工学）

関連URL

NECの研究開発 R＆D：データマイニング:
http://www.nec.co.jp/rd/datamining/

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報(日本語)　　NEC Technical Journal(英語)

Vol.65 No.2　ビッグデータ活用を支える基盤技術・ソリューション特集

ビッグデータ活用を支える基盤技術・ソリューション特集によせて
ビッグデータを価値に変えるNECのITインフラ

◇ 特集論文

データ管理 / 処理基盤

超高速データ分析プラットフォーム「InfoFrame DWH Appliance」
SDN技術で通信フローを制御する「UNIVERGE PFシリーズ」
大量データをリアルタイムに処理する「InfoFrame Table Access Method」
大量データを高速に処理する「InfoFrame DataBooster」
ビッグデータの活用に最適なスケールアウト型新データベース「InfoFrame Relational Store」
高い信頼性と拡張性を実現したExpress5800/スケーラブルHAサーバ
大規模データ処理に対するOSS Hadoopの活用
大容量・高信頼グリッドストレージiStorage HSシリーズ(HYDRAstor)

データ分析基盤

ファイルサーバのデータ整理・活用を支援する「Information Assessment System」
超大規模バイオメトリック認証システムとその実現
WebSAMの分析技術と応用例～インバリアント分析の特長と適用領域～

データ収集基盤

スマートな社会を実現するM2Mとビッグデータ
微小な振動を検知する超高感度振動センサ技術開発とその応用

ビッグデータ処理を支える先進技術

多次元範囲検索を可能とするキーバリューストア「MD-HBase」
高倍率・高精細を実現する事例ベースの学習型超解像方式
ビッグデータ活用のためのテキスト分析技術
ビッグデータ時代の最先端データマイニング
ジオタグ付きデータをクラウドでスケーラブルに処理するジオフェンシングシステム
柔軟性と高性能を備えたビッグデータ・ストリーム分析プラットフォーム「Blockmon」とその使用事例

◇ 普通論文

地デジTVを活用した「まちづくりコミュニティ形成支援システム」

◇ NEC Information

NEWS

スケールアウト型新データベース「InfoFrame Relational Store」が2つの賞を受賞

Vol.65 No.2
(2012年9月)

特集TOP