フジサンケイビジネスアイ賞

異種混合学習技術とビッグデータ分析ソリューションの研究開発

1. 日本電気株式会社情報ナレッジ研究所
2. 日本電気株式会社ビッグデータ戦略本部

藤巻 遼平[1]　森永 聡[1]　江藤 力[1]　本橋 洋介[2]
菅野 亨太[2]

1．緒　言

エネルギー・水・食料の需給を予測して限りある資源を効率的に利用したい、インフラの劣化を把握して計画的に補修したい、災害を予測して事前に対策をうちたい、商品の需要を予測して欠品や廃棄を減らしたい。これらの社会課題に対し、従来のような人間の経験と勘に頼る手作業の予測では、予測モデルの作成に長い時間がかかる、精度があがらない、大規模な予測ができないなどの問題があった。今、実社会から刻々と発生しているデータ、すなわち「ビッグデータ」が注目されている。大量に蓄積された様々なビッグデータの分析と利活用により、今まで気づかれなかった新しい科学的発見による知的価値の創造や、新しく獲得した知識の活用による社会的・経済的価値の創造やサービスの向上などが期待されている。ビッグデータの活用は、様々な社会課題を解決し、私たちの社会を安心、安全、効率、公平で豊かにする可能性を秘めている。市場もビッグデータの可能性に大きな期待を寄せており、ビッグデータ関連市場規模は加速的に成長を続けると考えられ[1]、日本でもビッグデータを活用した10兆円規模の新市場創出を総務省、経済産業省、文部科学省が国家をあげて取り組んでいる[2]。

ビッグデータ活用の中で、特に科学的根拠に基づいた将来予測をする「予測分析技術」への期待が高まっている。ビッグデータから予測モデルを作る作業の一部は、機械学習による高度化が進んでいるが、その取り扱いには高度なスキルが要求される。今のところ、データから実運用に耐えうる予測モデルを作るためには、数理的素養をもったデータサイエンティストによる試行錯誤が不可欠である。しかし、米国では2018年に14～19万人のデータサイエンティストが不足すると試算されるなど、データサイエンティストの圧倒的な不足が世界的な課題となっている[3]。日本では文部科学省を中心にデータサイエンティストの教育に力を入れているが[4]、加速的に増え続けるニーズに追い付いていないのが実情である。IoT（Internet of Things）の広がりとともに、実世界で大量のデータがリアルタイムに収集され、その迅速な利活用が求められつつある中、データサイエンティストの不足と属人的スキルへの依存という問題を、IC T 技術で解決することが大いに期待されている。

我々は、データサイエンティストが行ってきた試行錯誤を自動化する「異種混合学習技術」を開発・実用化した。データサイエンティストによる試行錯誤を自動化する試みは世界的にも類をみず、この実現に向けた努力は因子化漸近ベイズ推論という新しい機械学習理論を生み出した。我々は、電力や水の効率運用、食料廃棄の削減や在庫管理の効率化などを実現する予測ソリューション（以下、予測 SL）を実用化し、よりよい社会の実現に向けてさらなる研究開発を進めている。

2．開発の背景

検索、電子商取引、ソーシャルメディアなどのウェブサービスを中心に、我々は日々の生活でビッグデータに基づく予測分析を利用している。今後、ビッグデータ利用は、M2M や IoT といった実社会におけるデータ収集基盤の進歩と共に、交通渋滞、医療の充実や犯罪抑止といった社会的課題の解決や、電力網、水道網などの社会インフラの効率的運用、小売店舗管理や在庫管理をはじめとするビジネスの効率化などへ広がる事が期待され、これによって少なくとも10兆円規模の付加価値創出及び12～15兆円規模の社会的コスト削減効果があ

ると推定されている[2]。

予測分析の実社会への広がりと共に、「予測の透明性・公平性・説明責任」という新たな課題が生じている[5-7]。例えば、人命にかかわる医療や、ライフラインを支える社会インフラ、行政による社会保障など、実社会における予測分析では、予測の根拠を利用者へわかりやすく説明できることが求められている。複雑な非線形予測は、予測精度が高かったとしても、挙動がブラックボックス化されてしまう。一方で、線形回帰や決定木などは、単純でわかりやすい反面、複雑なビッグデータの挙動を捉える事が出来ず、予測精度が低くなってしまう。

精度とわかりやすさを両立するために、これまではデータサイエンティストが、規則性が切り替わる要因を想定し、その単位にデータを分割して、それぞれに線形回帰モデルのような単純なモデルを適用するという試行錯誤が広く行われていた(図1)。コンビニエンスストアにおけるおにぎりの売上予測を例にすると、平日はビジネスマンの購入が多く昼食時の商品陳列数と売上が高い相関を持つが、休日は家族連れが多くライバル店との価格差が売上と高い相関を持つ、といった具合に、シンプルな切り替えルールとパターンに応じて説明変数[1]を組合せる事で高い精度で予測できる。しかし、データの場合分けと説明変数の組合せのパターンは無限に存在し、その中からしらみつぶしにモデルを探すことは現実的ではない。一方で、この試行錯誤がデータサイエンティストの手作業に依存している限り、予測分析の実社会への普及に向けたボトルネックとなる。

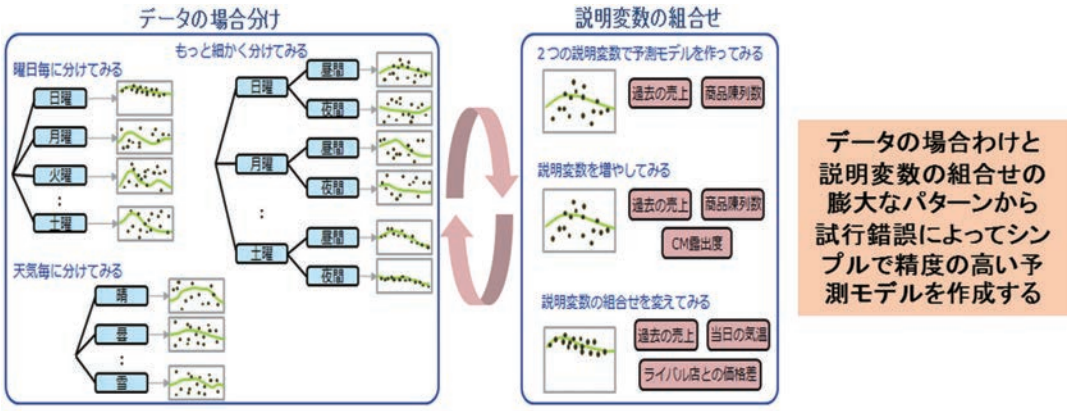

図1　データサイエンティストによる試行錯誤

我々は、このような背景から、データサイエンティストの経験に基づく手作業の試行錯誤を自動化し、社会のニーズにあった分析スピードで、予測モデルを作成する技術の開発を目指した。これによって、安心・安全・高精度な予測モデルを素早く作成し、予測 SL を実社会で広く普及させる事ができると考えた。

3. 異種混合学習技術

3.1　異種混合予測モデル

我々が、最初に取り組んだのが、精度が高く、かつ、人間が理解可能な予測モデルの開発である。サポートベクトルマシン[8]など非線形モデルと同じくらい予測精度がよく、線形

[1] 予測対象の変数(図1では売上)と高い相関を持つ予測にとって重要な変数の事

回帰のように利用者が予測の根拠を理解する事ができる。この相反する要求を検討するにあたり、我々は、複雑にみえるデータの挙動も、丁寧に解きほぐせば、比較的単純な挙動を組合せることで高い精度で予測できるという、データサイエンティストの現場のプラクティスを数学的に表現する事に挑戦した。そして、入力データを決定木形式のルールによって場合分けし、各場合で異なる説明変数を組合せた線形モデルで予測するモデルを開発した(図2)。異なる説明変数の組合せ(異種)による予測モデルを、組合せて(混合)予測するため、我々はこのモデルを異種混合予測モデルと命名した。

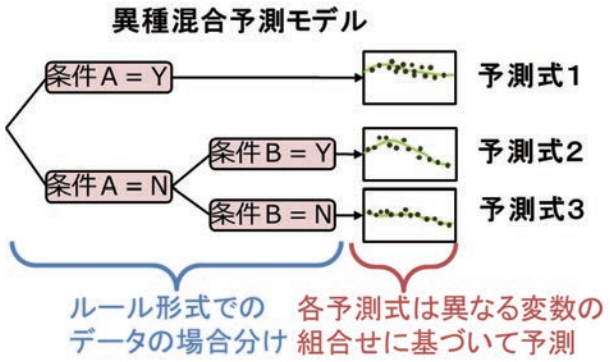

図2　異種混合予測モデル(Y と N はそれぞれ条件が成立する・しない事をあらわす)

異種混合予測モデルをデータから自動的に学習する事ができれば、データサイエンティストの試行錯誤を自動化する事ができる。しかし、異種混合予測モデルは、一見、単純であるが、複雑に絡まりあった複数の組合せ(予測式の組合せ、説明変数の組合せ、データの場合分け)によって膨大な数のモデル候補があり、そこから実用的な計算時間で探索を完了する必要がある。さらに、この探索には、①予測式の数を決める問題、②各予測式に利用される説明変数の数と種類(組合せ)を決める問題、③データを分割するルール構造を決める問題、という3つの「モデル選択問題」を同時に解く必要がある。モデル選択問題は、それ自体が難しい機械学習問題であり、それを3つ同時に解くという事は前代未聞の難問であった。

3.2　因子化情報量基準と因子化漸近ベイズ推論

我々は、以下の2つの技術的ポイントによって、①から③の3つのモデル選択問題を同時に解決し、異種混合予測モデルを自動的に学習する技術を開発した。

【因子化情報量基準】

機械が最適な異種混合予測モデルを探索するためには、探索の指針となる「モデルのよさ」を測る基準が必要である。予測モデルのよさを測る基準は、赤池情報量基準[9]、ベイズ情報量基準[10]など、統計学の分野で古くから研究されている。これらは、線形回帰のように単一のモデルで予測をする「正則モデル」に対してのみ適用可能である。しかし、異種混合予測モデルは複数のモデルを切り替えながら予測をする「特異モデル」であり、従来の情報量基準ではよさを正しく測る事ができない[11]　(図3)。特異モデルの情報量基準は、機械学習の重要な未解決問題であり、異種混合予測モデルの学習自動化は、新しい機械学習理論への

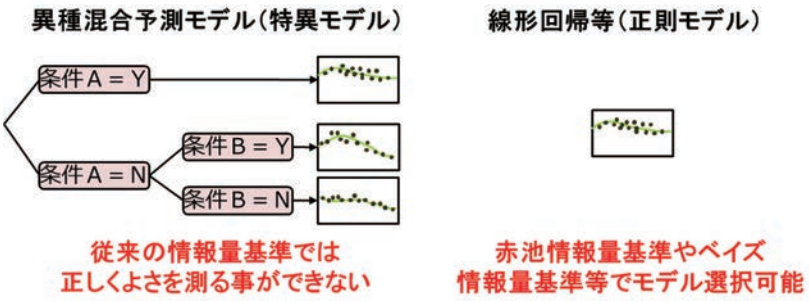

図3　異種混合予測モデルは複数のモデルが絡み合い従来の情報量基準が適用できない

挑戦でもあった。

我々は、図3から観察されるように、異種混合予測モデルは、複数の正則モデルが組み合わさって構成されており、この複数のモデルの絡まりによって従来の理論が適用できない点に着目した。そして、この「複数のモデルの絡まり」を数学的なテクニックを駆使して解消し、特異モデルのための新しい情報量基準を導出する事に成功し、その基準を因子化情報量基準と命名した。因子化情報量基準は、予測精度が低い、あるいはデータの場合分けや説明変数の組合せが複雑なモデルは値が低い。逆に、シンプルで予測精度の高いモデルは値が高く、データサイエンティストの経験的な基準とよく合致する。

【因子化漸近ベイズ推論】

機械が因子化情報量基準を最大化するモデルを見つけ出すためには、探索アルゴリズムが必要である。従来の正則モデルの探索は、各モデル候補に対して、まずはパラメータを最適化し、それぞれの情報量基準を算出し、その値を比較する事で候補から最もよいモデルを選ぶ。しかし、異種混合予測モデルは①から③で説明した複数の組合せ問題が絡まりあい、ほぼ無限にモデル候補が存在するため、それらの因子化情報量基準の値を一つ一つ算出し、しらみつぶしに検証する事はできない。一方で、膨大なモデル候補の多くは「筋が悪い」(予測精度が低い、モデルが複雑すぎて人間には理解できない)ため、データサイエンティストはそのようなモデルを試行錯誤から除外しているという事が経験的にわかっていた。筋の悪いモデルを探索から除外する事ができれば、最適な異種混合予測モデルを素早く探索できる。そこで、我々は図4に示される、3つのステップを順番に繰り返し解く事で、因子化情報量基準を最大にするモデルを探索するアルゴリズムを開発し、因子化漸近ベイズ推論と命名した。

各ステップは、①から③に対応する通常のモデル選択問題を個別に解けばよく、効率的に解くことができる。さらに、我々は各ステップにおいて、必ず因子化情報量基準の値を改善するモデルを見つける事ができる手順を発見した。これによって、「筋の悪い」(即ち因子化情報量基準の値が低い)候補を探索すること無く、高速に最適な異種混合予測モデルを見つける事が可能となった。この手順は、まさに図1で示したデータサイエンティストの試行錯誤を数学的に表現し、アルゴリズムとして実装したものである。我々は、因子化情報量基準・因子化漸近ベイズ推論による異種混合予測モデルの学習技術を異種混合学習と命名した。

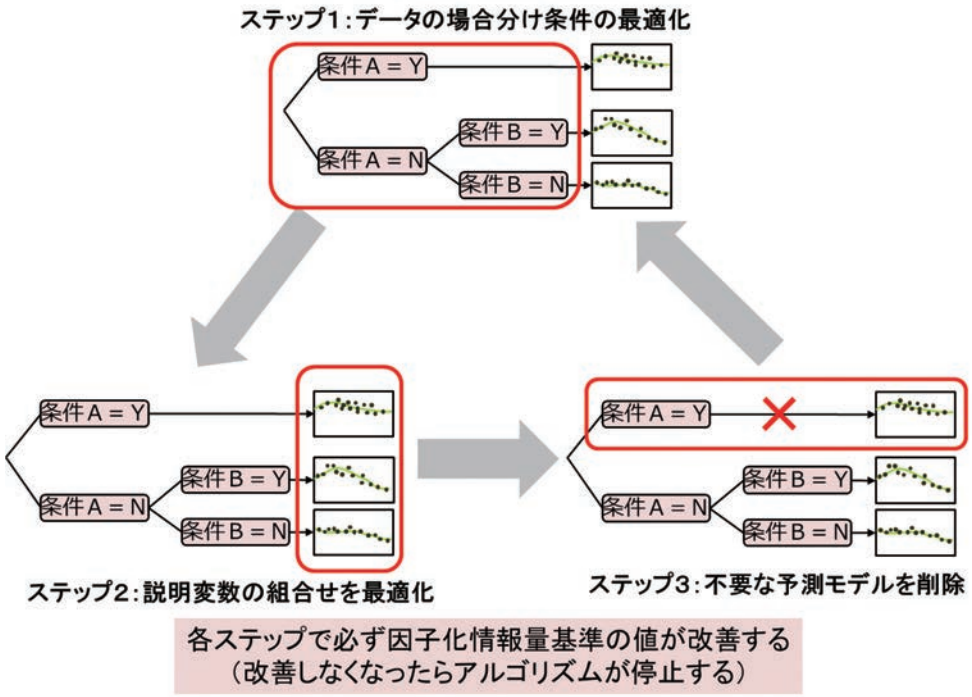

図4　因子化漸近ベイズ推論による異種混合予測モデルの最適化ステップ

3.3　ベンチマークデータにおける性能評価

我々は、異種混合学習の効果を確認するために、最初に、モデル選択性能の検証実験を行った。シミュレーションによって「真のモデル」を作成し、そのモデルからデータをサンプリングし、学習データを入力として異種混合学習し、「真のモデル」を復元できるかを評価する。まず、我々は図5に示される「真のモデル」を作成した。4つのデータの場合分け条件（四角形）と5つの予測式（円形）から成り（左図）、また各予測式は11の説明変数のうちランダムに選ばれた少数の変数のみを用いて予測を行なう（右図）。この真のモデルからデータを生成し、探索する予測式の最大数を32に設定し、因子化漸近ベイズ推論によってモデルの探索を行なった。この実験では、考えうるモデルの候補数が、$(2^{11})32 = 9.17 \times 10^{105}$という膨大な数となり、しらみつぶし的な方法では、スーパーコンピューターを使ったとしても探索に非実用的な時間がかかってしまう。一方で、因子化漸近ベイズ推論を使った探索ではたったの18回のモデルの探索（図6の上図。1回の探索は図4の3ステップからなる）によって因子化情報量基準を最大化し、真のモデルとほぼ等価なモデルを発見する事に成功した（図6）。この学習にかかった計算時間は、汎用計算機でたったの10秒という驚異的かつ実用的にも十分な速度であった。これによって、①から③で述べた3つのモデル選択問題を自動的・高速・正確に解決できる事を、数値的に示すことができた。

次に、予測精度を確認するために、異種混合学習の他に、線形回帰、決定木、サポートベクトルマシンを比較対象として、UCI データレポジトリ [12] の複数の標準データセットに対するベンチマーク評価を行った。図7に示されるように、異種混合学習は、線形回帰や決定木を大きく凌駕し、目標とした非線形予測器のサポートベクトルマシンと同程度の予測精

度を達成した。

以上の実験によって、異種混合学習によって、確かにデータサイエンティストの手を煩わせる事なく、高精度かつ高解釈の異種混合予測モデルを、自動的に素早くデータから学習できる事が確認できた。因子化情報量基準、因子化漸近ベイズ推論、そして異種混合学習に関する論文は、データマイニング及び機械学習の最高峰会議へ、2011年から2014年の4年連続で合計8本の論文が採録され[13-19]、予測分析の自動化さらには特異モデルの学習理論として、学術的にも世界で高く評価されている。

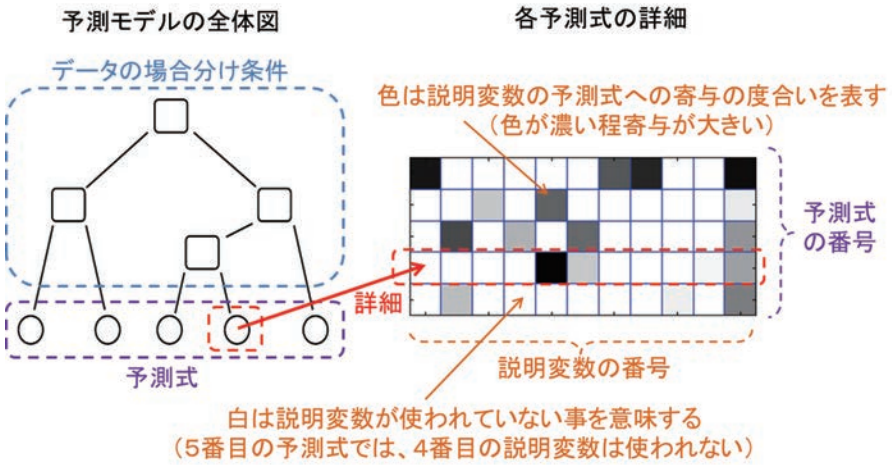

図5　モデル選択性能の検証実験で作成したシミュレーションモデル

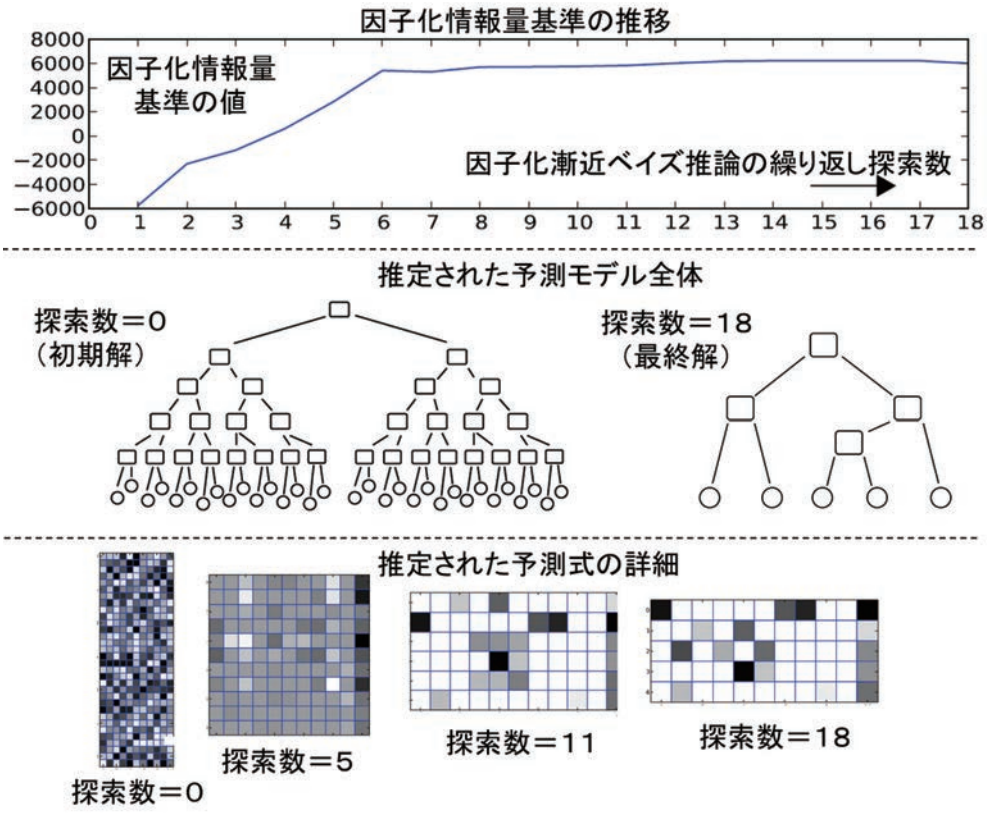

図6　異種混合学習の学習プロセスのシミュレーション結果

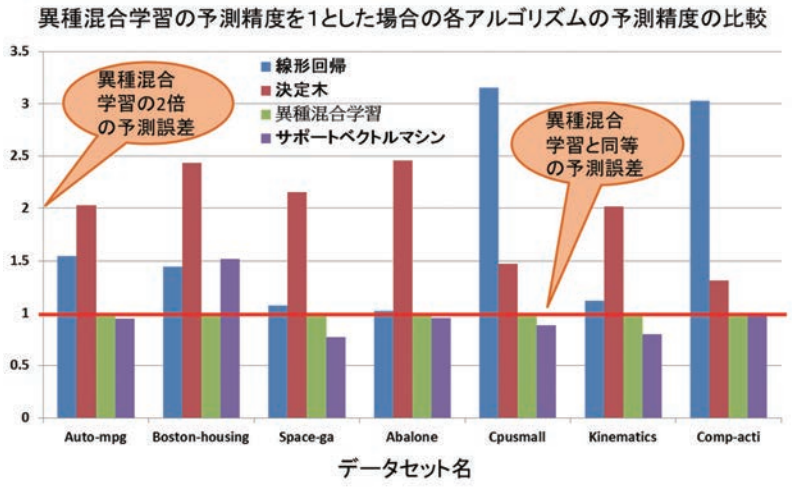

図7　ベンチマークデータでの予測精度の比較

4．異種混合予測ソリューション

我々は、異種混合学習技術を用いた様々な予測 SL の開発を進め、既に複数の商用化実績を持つ。以下では、代表的な具体例を紹介する。

4.1　エネルギー需要予測 SL[20]

電力や熱などのエネルギー需要の正確な予測は、発電、蓄電や蓄熱の調整、冷暖房の制御など、エネルギーを効率的に運用する様々な仕組の根幹となる基本技術であり、従来から自己回帰モデルなどの時系列分析手法に基づく方法などが研究されている。しかし、例えば、オフィスビルと住宅、平日と休日、昼間と夜間、天気や気温といった様々な要因によって需要のパターンは大きく異なり、従来の手法によって高い予測精度を達成するためには、ビルの特性に合わせた予測モデルの調整を人手で行う必要があり、都市の建物単位の需要を高精度に予測することは難しかった。

NEC は大林組と、エネルギー需要予測及びそれを活用したビル群のエネルギー管理のスマート化に向けたエネルギー需要予測 SL の実証を行なった。大林組技術研究所で収集された、過去2年間の電力使用量、空調に用いた熱量(温水熱量 / 冷水熱量)、気象、営業日、日付、在籍者数などの各種データを基に、将来の電力使用量および熱量を予測した。その結果、「冬期営業日の昼間」、「夜間」、「祭日」などで異なる規則性を自動的に発見し、24時間後や1ヵ月後などの電力使用量・熱量を、人手による複雑なデータ分割作業を行うことなく、高精度に予測する事ができた。自己回帰モデルとの比較では、自己回帰モデルが平均誤差率27.1％に対して、異種混合学習は16.8％という予測精度を達成した。また、異種混合学習によって算出された予測モデルは、大林組の現場でエネルギーのスマート化を推進する現場の経験とも合致するものであった(図8)。この結果、異種混合学習を利用したエネルギー需要予測 SL は、大林組の技術研究所内のすべてのビルを対象にしたエネルギースマート化プロジェクトのキーコンポーネントとして採用され、エネルギー運用のスマート化により約20％の電力使用量削減につながる事が期待されている(図9はエネルギー需要予測 SL の画面例)。

世界的にエネルギー需要は増え続けており、また日本のエネルギー自給率が5％前後と低く、エネルギーを如何に効率的に運用するかは、個人や一企業のエネルギーコスト低減を超えて、都市スケールあるいは国スケールでの効率化が重要となっている。異種混合学習に基づくエネルギー需要予測 SL(図10)によって、数千～数万棟のビル・建物のエネルギー需要の高精度予測を自動化する事が可能となり、個別ビルの電力消費・調達の最適化のみならず地域や社会全体での、よりよいエネルギー運用の実現に貢献していく。

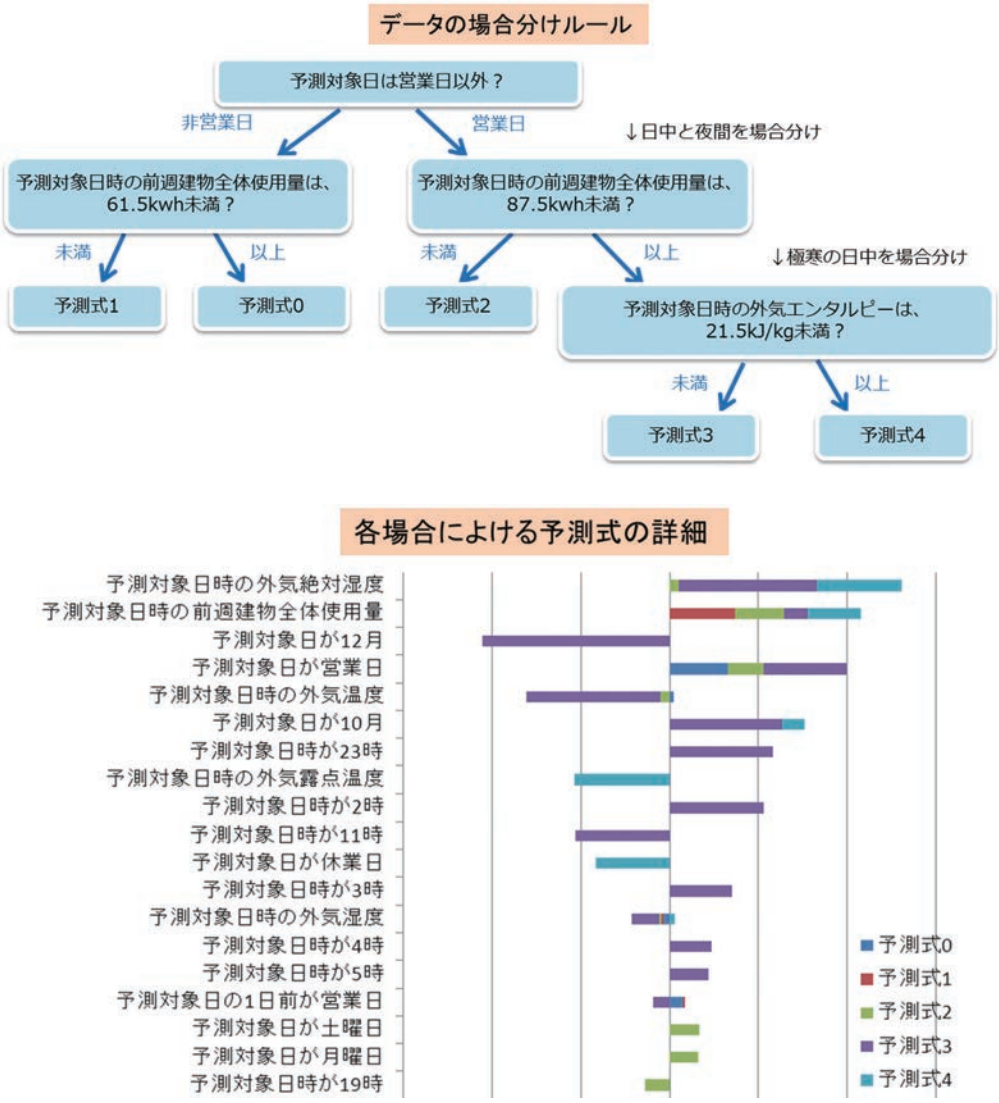

図8　エネルギー需要予測 SL の実証実験で得られた異種混合予測モデル

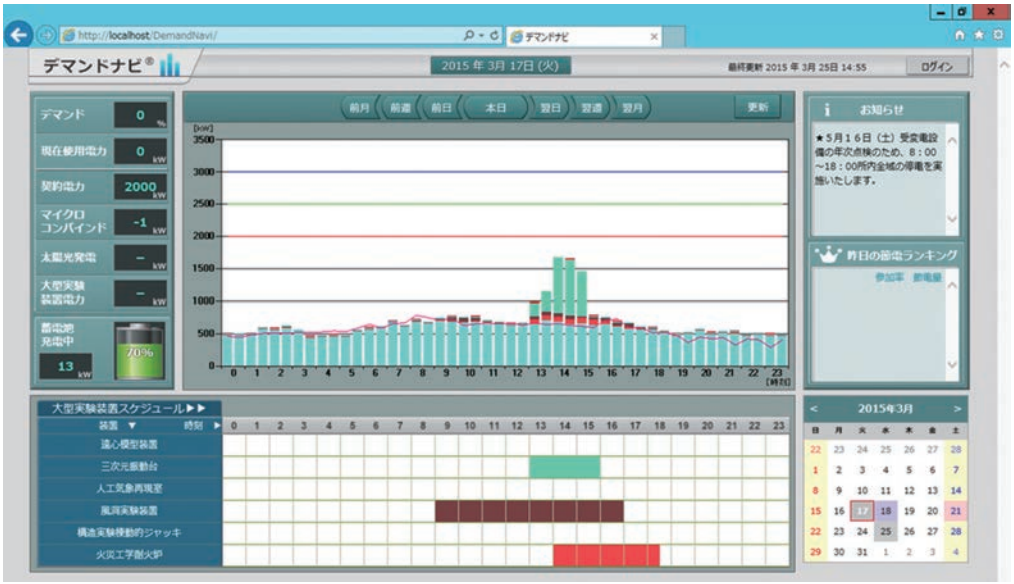

図9　エネルギー需要予測 SL のシステム画面

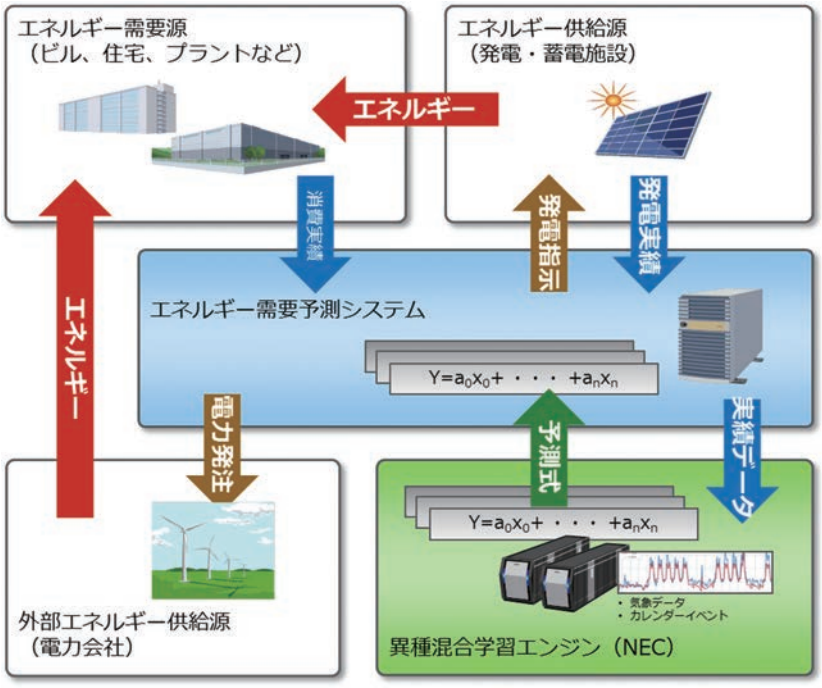

図10　エネルギー需要予測 SL の模式図

4.2　広がる異種混合予測ソリューション

異種混合予測 SL は、エネルギー需要予測のみならず、様々な事例において実用化され、また広がっている。

【需要予測型自動発注 SL[21]】

小売店舗向けに商品の需要予測型自動発注 SL を開発した（図11）。天候や気温、過去の売上傾向、イベントや休日などの様々な情報を異種混合学習により分析することで、これまで不可能であった数千店舗×数千商品の大規模、かつ、精緻な需要予測を行い、最適な発注量・陳列量の算出を自動化した。実証実験では、売れ残りによる廃棄を人手で発注する場合と比較して30％削減できた。また、欠品を大幅に低減し、お客さまに満足行く店舗作りに役立った。さらに、我々はこのソリューションを新商品の売上予測へ応用し、新商品の需給調整や新商品開発の最適化を実現する予測 SL の開発へ取り組んでおり、実用化目前である[22]。

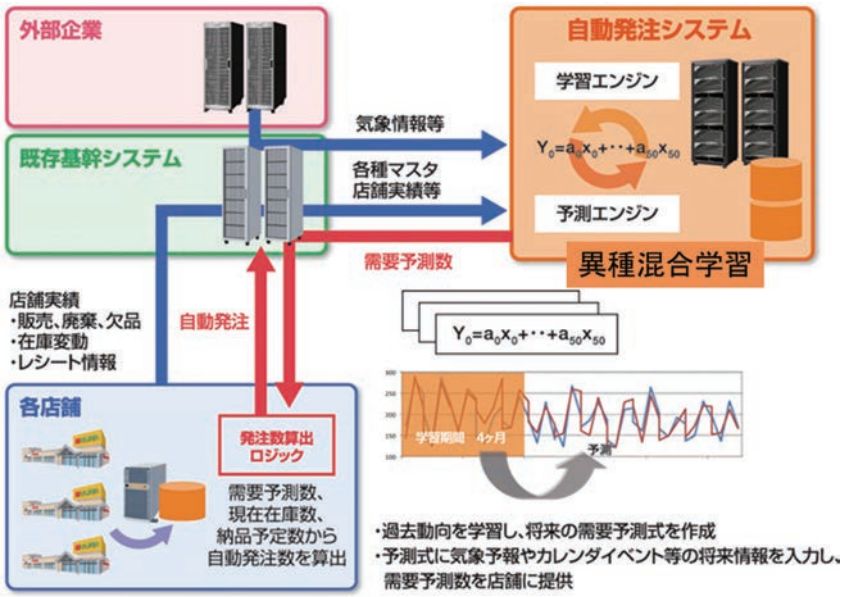

図11　需要予測型自動発注 SL の模式図

【水需要予測 SL】

水資源の効率的な利用や運用の高度化に向けては、水需要予測 SL を開発した。都市や地域に大量に存在する需要点（各地区など）における水需要傾向を異種混合学習によって自動的にモデル化し予測する事で、浄水・配水・貯水の計画を精緻化・最適化したり、想定需要と実際の使用量から配水パイプからの漏水を発見するなど、水の安定供給や水損失の低減・水資源の有効活用に貢献している。

【補修部品需要予測 SL[23]】

在庫管理の高度化に向けては、補修部品の需要予測 SL を開発した。NEC フィールディングが保有する約1万点の出荷頻度の高い部品の需要を予測し、在庫量を適正化する事に

よって、これらの在庫保有量が約2割削減する事が確認された。これは保管コストや保管スペースの削減といった効果もある。現在、補修部品需要予測SLはNECフィールディングで運用が開始され、また2015年度より製造業向けの提供を予定している。

5．結　言

ビッグデータから、高精度かつ高解釈の予測モデルを自動的に学習する異種混合学習技術を開発し、異種混合予測ソリューションを実用化した。これまでのデータサイエンティストによる試行錯誤を自動化し、実用的なスピードで予測モデルを作成できるようになり、エネルギーや水の効率的利用などの社会課題や、店舗や在庫の管理などのビジネス課題の解決で実績がある。様々な課題に直面する現代社会において、その解決に資する予測ソリューションへの期待は大きい。我々は、今後もビッグデータ活用技術の開発に務め、その成果を実社会に適用することで、安心、安全、効率、公平で豊かな社会の実現に貢献する所存である。

参考文献

［1］ IDC「Wo rldwide Business Analytics Technology and Services 2013-2017 Forecast」

［2］ 平成24年版総務省情報通信白書
http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/pdf/n2010000.pdf

［3］ Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.

［4］ 文部科学省委託事業「データサイエンティスト育成ネットワークの形成」平成25年度事業報告書、http://datascientist.ism.ac.jp/pdf/H25DSTN.pdf.

［5］ Fairness, Acountability, and Transparency in Machine Learning,
http://www.fatml.org/index.html

［6］ Cynthia Rudin, Algorithm for interpretable machine learning, Invited Talk in 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining（KDD）, 2014

［7］ Big Data: Seizing Opportunities, Preserving Values, Exective Office of the President（White House）, 2014,
https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

［8］ Vladimir N. Vapnik, The nature of statistical Learning Theory, Springer-Verlag New York, 1995.

［9］ Akaike, H., Information theory and an extension of the maximum likelihood principle Proceedings of the 2nd International Symposium on Information Theory, 267-281, 1973

［10］G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6（2）:461.464, 1978.

［11］S. Watanabe, Algebraic Geometry and Statistical Learning Theory, Cambridge University Press, 2009

［12］M. Lichman, UCI Machine Learning Repository, http://archive.ics.uci.edu/mi, 2013.

［13］Ryohei Fujimaki, Yasuhiro Sogawa, Satoshi Morinaga: Online heterogeneous mixture modeling with marginal and copula selection. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining（KDD）, 2011
［14］Ryohei Fujimaki, Satoshi Morinaga: Factorized Asymptotic Bayesian Inference for Mixture Modeling. Proceedings of the The fifteenth international conference on Artificial Intelligence and Statistics（AISTATS）, 2012
［15］Ryohei Fujimaki, Kohei Hayashi: Factorized Asymptotic Bayesian Hidden Markov Model. Proceedings of the 25th international conference on machine learning（ICML）, 2012
［16］K. Hayashi and R. Fujimaki, "Factorized Asymptotic Bayesian Inference for Latent Feature Models", 27th Annual Conference on Neural Information Processing Systems（NIPS）, 2013.
［17］Riki Eto, Ryohei Fujimaki, Satoshi Morinaga, Hiroshi Tamano, Fully-Automatic Bayesian Piece-wise Sparse Linear Models, Proceedings of the 17th International Conference on Artificial Intelligence and Statistics（AISTATS）, 2014
［18］Ji Liu, Ryohei Fujimaki and Jieping Ye, "Forward-Backward Greedy Algorithms for General Convex Smooth Functions over A Cardinality Constraint", Proceedings of the 27th international conference on machine learning（ICML）, 2014
［19］H. Oiwa and R. Fujimaki, "Partition-wise Linear Models", 28th Annual Conference on Neural Information Processing Systems（NIPS）, 2014.
［20］NEC プレスリリース（2013年10月29日）、「大林組と NEC、ビッグデータ分析技術を活用してビルのエネルギー需要を予測する実証実験を共同実施」http://jpn.nec.com/press/201310/20131029_02.html
［21］NEC ホームページ - 製品／サービス価値向上・改善
http://jpn.nec.com/bigdata/example/value.html
「異種混合学習技術」を用いたビッグデータのソリューションサービス（需要予測ソリューション）の紹介。
［22］2015年2月16日掲載日経産業新聞（7面）「酒類売れ筋　データで予測　アサヒビールとNEC」
［23］NEC プレスリリース（2014年11月12日）、「NEC と NEC フィールディング、ビッグデータ分析技術を活用して補修用部品の需要を予測」
http://jpn.nec.com/press/201411/20141112_02.html