特集号
招待論文

分析プロセス自動化・標準化への挑戦―実践に基づく考察―

藤巻 遼平[†1]　本橋 洋介[†2]

[†1] NEC Knowledge Discovery Research Laboratories　[†2] 日本電気（株）

ビッグデータから価値のある法則を導き，実際の業務で活用するためには，単なる統計・機械学習ツールの実行だけでなく，分析目的の設計，データの前処理，モデルの設計，業務フローへの組み込み，といった一連のプロセスが必要である．この「分析プロセス」を実行するデータサイエンティストには，顧客の業務と分析の両面の理解が求められ，現状では個人のスキルに依存しているケースがほとんどである．本稿では，筆者らが第一線のデータサイエンティストとして挑戦する分析プロセスの自動化・標準化への取り組みを説明する．

1. はじめに

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講じたい，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，精度が上がらない，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．大量に蓄積されたさまざま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．

ビッグデータから価値のある法則を導き業務で活用するためには，分析目的の設計，データの前処理，モデルの設計，業務フローへの組込み，といった一連のプロセスが必要である．いわゆるデータサイエンティストの業務に明確な定義はないが，単なる数理統計や機械学習ツールの知識や実行だけでなく，顧客（エンドユーザ）の業務に対する正しい理解と顧客との対話を含む試行錯誤的かつ高度な分析作業が求められ，現状では個人のスキルに依存しているケースがほとんどである．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[1]．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[2]，加速的に増え続けるニーズに追い付けていないのが実情である．

IoT（Internet of Things）の広がりを始め，ビッグデータ分析による社会や顧客課題の解決はますます期待とニーズが高まっている．その可能性を広げるために，ICTによってデータサイエンティストによる分析プロセスの自動化，標準化が求められている．本稿では，筆者らが第一線のデータサイエンティスト，そして分析技術研究者として挑戦する，分析プロセス自動化・標準化に向けた取り組みを説明する．

2. 分析プロセスの全体像と課題

1996年にUsama Fayyadらは，従来はメインのアルゴリズムのみに焦点が当たっていたデータ分析を，前処理やデータ変換，モデルのチューニングや後処理といった一連のプロセス，KDD（Knowledge Discovery and Data minng）プロセス，として定義した[3]．そして，現在では，データ収集，目的設定，そして分析から得られるモデルの運用を含めた広い範囲を，分析プロセスとして捉えることができる（図1）．

データサイエンティストには，分析のフェーズによってさまざまな役割が求められる．目的設定・データ収集のフェーズ（図1の左）では，顧客の保有するデータと業務を理解し，顧客の課題を明らかにし，それを解決するソリューション（SL）を分析問題へブレイクダウンし，それによって得られる価値の定量化や評価の基準を，顧客との対話を通じて設計・提案する．分析のフェーズ（図1のKDD Process）では，データを適切に前処理し，

特徴空間やモデルを設計し，得られた結果が顧客の課題を解決に資するかを顧客と共に検討する．そして，運用のフェーズでは，得られた分析モデルを顧客の業務システムへ組込むためのシステムを設計し，構築し，運用する．

本稿では，筆者らが取り組む分析プロセスの自動化，標準化について紹介する．第3章では，分析プロセスを業務フローとして標準化する取り組みを紹介する．データサイエンティストには，上述したように，さまざまなフェーズで分析の専門家としての役割が求められる．業務フローとして標準化し，一部をテンプレート化することで，データサイエンティストの業務を均質化，効率化することを目指している．第4章と第5章では，特に予測分析に焦点を絞り，予測モデルと特徴空間の設計を自動化するための技術を紹介する．予測モデルと特徴空間の設計は，一般的にデータや分析技術に関する深い理解に基づく試行錯誤的な作業となり，運用に資する結果を得るためには時間と工数を要する．機械学習技術によって，また，試行錯誤を自動化することによって，高精度な予測モデルを素早く顧客へ提供し試行錯誤のプロセスを迅速化するとともに，大量の予測モデルを運用することを可能としている．

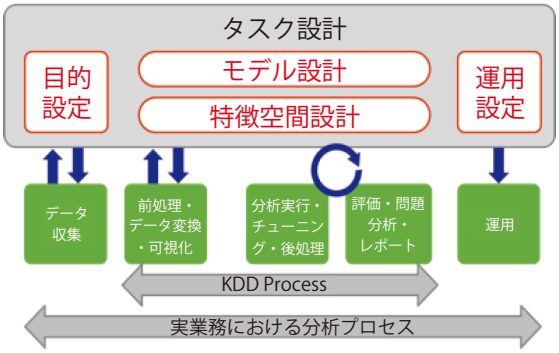

図1　分析プロセスの模式図

3.　分析プロセスの標準化への取り組み

データ分析業務および分析を伴うシステムの開発業務は，一般のシステム開発業務と異なる．図2に分析系システム開発の流れと一般のシステム開発の流れの違いを示す．分析系システムの開発は，要件定義の前に，データを用いた価値検討が行われる点や，システム開発後にシステムの要件定義のための分析が行われる点が，一般のシステムと異なる．特に，システム開発の要件定義用の分析とは，モデルの更新頻度やモデル作成に用いるデータの期間・種類などの検討をするための分析を指す．ほかに，たとえば商品の需要予測では，新商品が出たときの取り扱いをどうするかといった，データが変化した際の取り扱いについて決定することを含む．図2のような流れを実現するためにデータサイエンティストは，顧客へのヒアリング，分析プロジェクトの企画・提案，分析結果の解釈・説明，分析システム設計，運用サポートなど多岐にわたる役割・業務を担当する．そして，これらの業務の実施方法・手順は各人に任されており，成果物の質や業務の工数は作業者のスキルによってばらつきが大きくなっている．そこで，我々は，分析業務フローの標準化および分析成果物の定型化（テンプレート化）に取り組んでいる．以下，これらの取り組みを説明する．

3.1　分析業務フローの標準化

図3に，我々が策定した分析業務フローの全体図を示す（なお，分析の後にシステム開発を行う場合のフローを割愛している）．このように規定することで，各フェーズにおいて行わなくてはならない業務の抜けやもれを防止することができる．業務フローの設計は，これまで我々が経験したプロジェクトの流れを基に共通項を見出して実施した．たとえば秘密保持契約やデータ授受方針の合意については，通常のプロジェクトと比較しても特に重要に

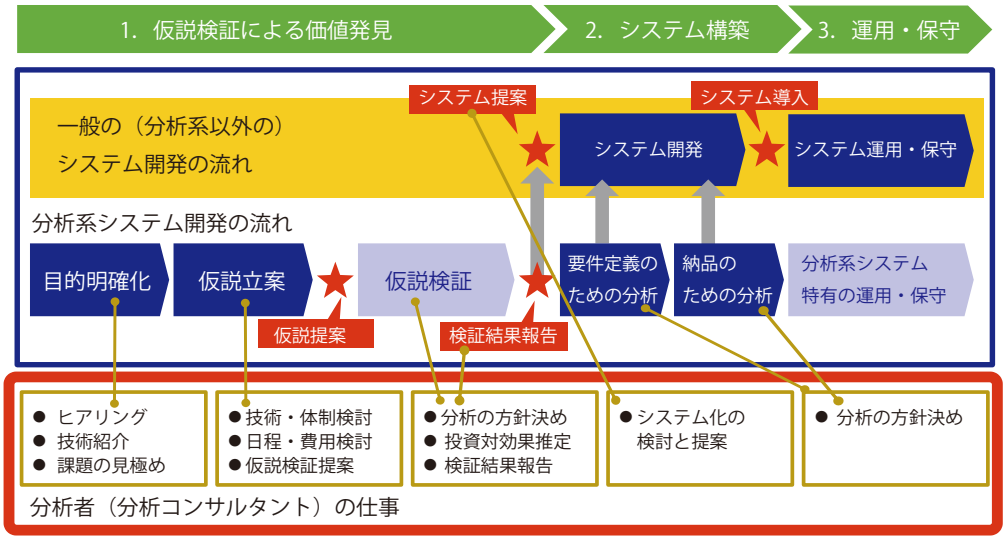

図2　分析業務プロジェクトの流れと分析者の役割

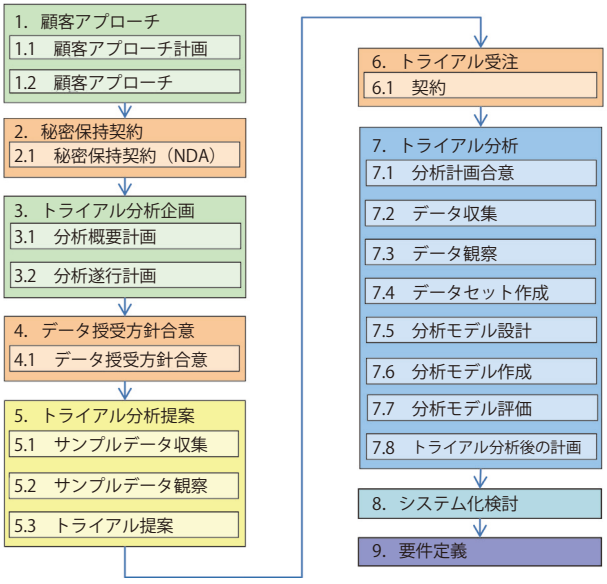

図 3　分析業務フロー全体

なることが多い．受け渡しの方法，終了後の廃棄手順，利用対象者などについてを決定することで後のトラブルを防止することが重要である．また，データ観察をフロー内に明記することによって，おろそかになりがちな分析前データ観察を必ず行うようにするなどの工夫を施している．

図3の業務フローは全体を示すために，大項目のみが記載されており，実際の業務フローはさらに細分化された項目によって構成されている．たとえばトライアル分析は前処理・分析処理・後処理がスパイラル的に何回も実施される形となる．分析結果が出た後に課題の整理・再分析方針を決定した後に，分析方針に従ってフローを巻き戻ることは分析プロジェクトでは必ず起こるものである．我々は，このような手戻りフローを明記することで，プロジェクト進捗中に，何サイクルの分析を行ったか，どのようなタイプの手戻りを頻繁に行っているかといったことを「可視化」し，その後のプロジェクトの業務効率化につなげている．

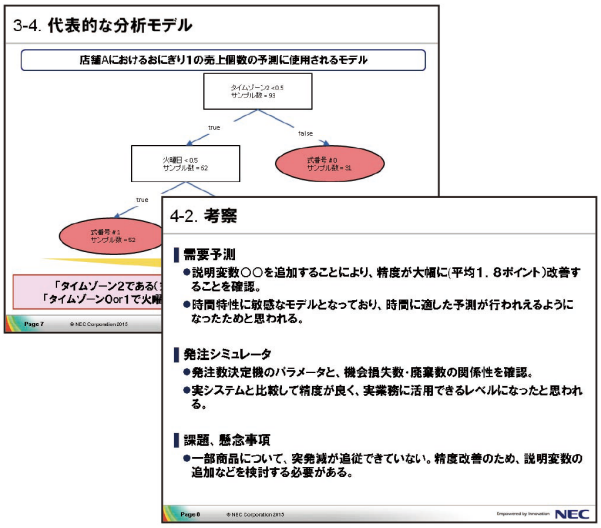

図 5　トライアル報告書テンプレートの例

3.2　分析成果物の定型化

前節で述べたフローの設計に続き，業務フロー各段階における成果物とその担当者を規定したWBS（Work Breakdown Structure）を作成した（図4に一部を抜粋表示）．このように成果物を規定することで，成果物の管理をしやすくし，過去の業務の成果物の再利用性を向上させる効果がある．

また，分析成果物の定型化（テンプレート化）においては，ヒアリングシート・提案書・報告書などの成果物の定型化を行っている．これは，顧客のビジネス要求を理解し，それを解決するための分析タスクを設計するスキルや，分析結果を解釈して知見化するスキルに個人差があることで提案書や報告書の質に大幅な差があるからである．図5に報告書テンプレートの例を示す．テンプレートには，図表などの例のほかに，提案や報告時に検討すべきポイントを記載している．これは，当初作成したテンプレートを運用しても成果物の均質化が実現できなかったため行っている施策である．たとえば機械学習で生成した予測モデルを解釈して，顧客の業務で活かし得る知見をまとめ顧客へ報告するレポートの質が当初のテンプレートでは安定しなかった．これは資料内の構成だけでは，低スキル者は何を埋めてよいかが分からないためで，過去の事例やそのときの考え方など，検討ポイントをガイドラインとして示すことで，成果の質の改善につなげようとしている．

				タスク入出力（★：分析標準テンプレートあり）		役割分担（●：作業者，□：支援者）※1人が複数の役割を担う場合がある				
L1：フェーズ	L2：アクティビティ	L3：タスク	L4：作業内容	INPUT（参照物）	OUTPUT（成果物）	お客様	営業	NEC（ドメイン E）	NEC（分析 C）	NEC（分析 E）
…	…	…	…	…	…	…	…	…	…	…
7. トライアル分析	7.1 分析計画合意	7.1.1 顧客課題整理	ヒアリング結果を 基に，顧客課題を整理する．	★お客様情報シート ★顧客アプローチ計画書	★トライアル計画書			□	●	
		7.1.2 プロジェクトゴール明確化	「NEC の保有技術によって，顧客課題に対する解決策が提供できることの検証」を，プロジェクトのゴールとして設定する．					□	●	
		7.1.3 顧客入手データ整理	顧客から，いつどのデータを入手できるかを整理する．					□	●	□
		7.1.4 分析 内容決定	分析目的に結びつくような分析内容を設定する．					□	□	●
		7.1.5 仮説立案	顧客課題解決につながる要因を，仮説として立案する．仮説に合ったデータが入手できているかを確認するため，「顧客入手データ整理」のタスクに戻ることがある．					□	□	●
		7.1.6 KPI 整理	顧客価値を探るため，KPI を整理する．					□	●	
		7.1.7 ROI 概算算出	顧客価値を金額で表現するため，ROI を試算する．					□	●	
…	…	…	…	…	…	…	…	…	…	…

図 4　分析 WBS の例（一部を抜粋）

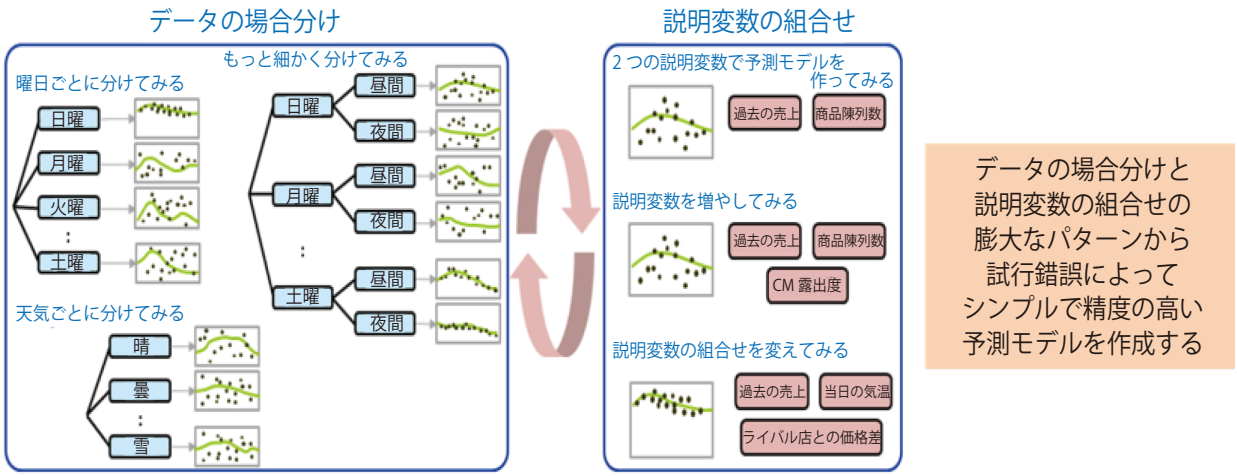

図 6　データサイエンティストによる試行錯誤を通じた予測モデルの設計

4.　予測モデル設計の自動化

4.1　予測モデルの設計と課題

ビッグデータ活用の中で，特に機械学習技術に基づいて将来予測をする「予測分析技術」への期待が高まっている．IoTといった実社会におけるデータ収集基盤の進歩とともに，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，予測分析は実社会への広がりを見せている．

これに伴って「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[4],[5],[6]．たとえば，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分かりやすく説明できることが必要となる．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，単純で分かりやすい反面，複雑なビッグデータの挙動を捉えることができず，予測精度が低くなってしまう．

ビッグデータから高精度な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，精度と分かりやすさを両立するために，通常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なわれている（図6）．コンビニエンスストアにおけるおにぎりの売上数予測を例にすると，平日はビジネスマンの購入が多いため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いため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を組み合わせることで高い精度で予測できる．また，企業には，さまざまなデータに対する分析課題が数多くあり，1つの分析テーマにかけられる時間は限られている．予測モデル設計の自動化によって，データサイエンティストの分析スピードを高め，少人数であっても多くの分析課題を解決することができる．

4.2　異種混合学習

本節では，筆者らが開発したデータサイエンティストによる予測モデルの設計（図1）を自動化するための異種混合学習技術を紹介する．異種混合学習の基本原理は，因子化漸近ベイズ推論という，筆者らが開発した最先端の機械学習理論に基づいており，詳細は参考文献[7],[8],[9],[10],[11],[12]などを参照．本稿では技術的詳細は割愛し，分析プロセス自動化の観点からポイントを説明する．

4.2.1　精度と分かりやすさの両立

異種混合学習では，精度と分かりやすさを両立するために，図7で示される「異種混合予測モデル」を用いる．

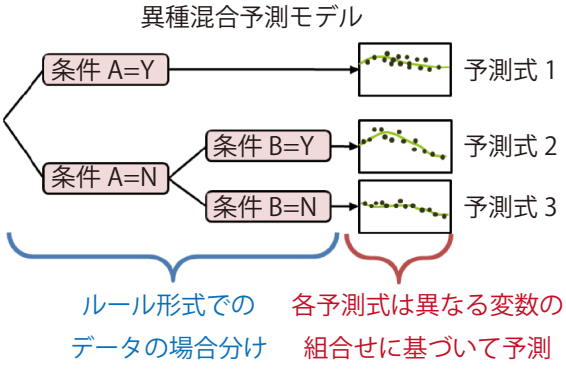

図 7　異種混合予測モデルの模式図

図6で示されるように，データサイエンティストは，データの場合分けと，線形モデル（重回帰分析やロジスティック回帰分析）を組み合わせることで，精度が高く説明力の高い予測を実現している．異種混合予測モデルは，同様に決定木形式のルールに基づくデータの場合分けと，各場合で異なる説明変数を用いた線形予測を組み合わせる．これによって，予測の根拠をブラックボックス化させることなく，どのような条件化で，どのような説明変数で予測すべきかを顧客へ説明することができる．

4.2.2　パラメータの自動決定

一般的に，予測分析アルゴリズムには，手で設定するパラメータが存在する．その最も代表的なものが，正則化パラメータである．学習データで予測精度が高くても，モデルが複雑になりすぎると，予測が学習データに過適合し，実際に運用する際には予測精度が大きく悪化する（過学習という）．そこで，学習データの予測精度とモデルの複雑性をバランスさせる「つまみ」を調整することが必要となる（図8）．正則化パラメータは，たとえば，決定木であれば木の深さあるいは木の深さと予測精度のバランス量，リッジ回帰分析ではL2正則化の強さである．正則化パラメータは，たとえば交差検定法とグリッドサーチを組み合わせることで機械に調整させることができるが，異種混合予測モデルでは，データの場合分けの複雑さと各予測式の複雑さが絡まっており，非現実的な計算量を必要とする．一方で，適切なバランスはデータによって異なり，手動によるチューニングには相応の経験と工数が必要となる．

異種混合学習では，因子化漸近ベイズ推論とよばれる独自の機械学習理論[7],[8]に基づいて，このような正則化パラメータは，データから自動的に調整されるため，最適な精度を出すためのチューニングに工数をとられることなく，学習されるモデルの解釈や運用方法の検討など，ほかの作業に注力をすることができる．

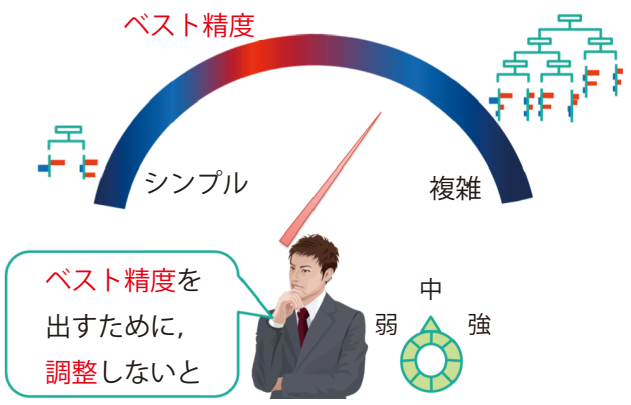

図8　精度と複雑性のバランスを調整

4.2.3　説明変数の組合せ最適化

4.1節のコンビニエンスストアのおにぎりの売上数予測で説明したように，正確な予測を行うためには，状況（データの場合分け）に応じて適切に説明変数を組み合わせる必要がある．しかし，どの変数を用いるかは，データや数理統計に関する知識だけでなく，業務に関する知識が必要となり，データサイエンティストが試行錯誤をするためには多大な工数を必要とする．交差検定法や，赤池情報量基準や，ベイズ情報量基準を用いれば，説明変数の組合せの候補から最も良いモデルを算出することができる．しかし，各候補に対してモデルを学習する必要があり，膨大な組合せの候補を調べることは難しいため，データサイエンティストが「筋のよい」候補を手で設定する必要がある．

一方で，異種混合学習では，説明変数の候補だけを入力すれば，各データの場合分けに対して，組み合わせる説明変数の数と種類を自動的に決定する．これによって，データサイエンティストは，説明変数の組合せの試行錯誤に時間をとられることなく，顧客との議論を通じた説明変数の候補の設計に注力することができる．また，手作業では調べることができなかった説明変数の組合せを調べることができるため，より精度の高い説明変数の組合せを見つけることができる．

4.3　エネルギー需要予測への適用

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さまざまな仕組み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，たとえば，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さまざま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．

日本電気（株）は（株）大林組と，エネルギー需要予測およ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カ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，高精度に予測することができた．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9**）．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，（株）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ことが期待されている．異種混合学習に基づくエネルギー需要予測SL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ことが可能となり，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，より良いエネルギー運用を目指している．

4.4　実践に基づく課題・考察

これまでに述べたプラクティスを実践することによって，分析フロー全体の流れ，作業項目，成果の標準化による業務の均質化（3章）や，予測モデリングの高度化，業務スピード向上（4章）について，複数の実証実験などを通じて有効であることが確認された．一方，複数の事例を通じて，いくつかの課題が明らかになった．

1つ目の課題は，図3の7.4（データセット作成）における説明変数の設計ノウハウの重要性である．異種混合学習は，「説明変数の候補集合」から自動的に説明変数の組合せを最適化することで，データサイエンティストによる試行錯誤を自動化した．一方で，さまざまなデータに対して説明変数の候補集合を準備することそのものが時間のかかる作業であり，データや対象ドメインに対するノウハウが必要となる．我々は，この課題に対して2つのアプローチを実施している．1つ目は，説明変数候補集合の設計自体を機械学習によって自動化するという試みである．2つ目は，類似の分析に関して説明変数セットを標準化し，ライブラリ化するというアプローチであり，これは特に「よくある分析に対して素早く結果を導出する」というシーンで効果を発揮する．これらの取り組みに関しては，詳細を5章で説明する．

2つ目の課題は，図3の7.7（分析モデル評価）における評価指標の標準化である．分析モデルの性能は，判別誤差やRMSE（Root Mean

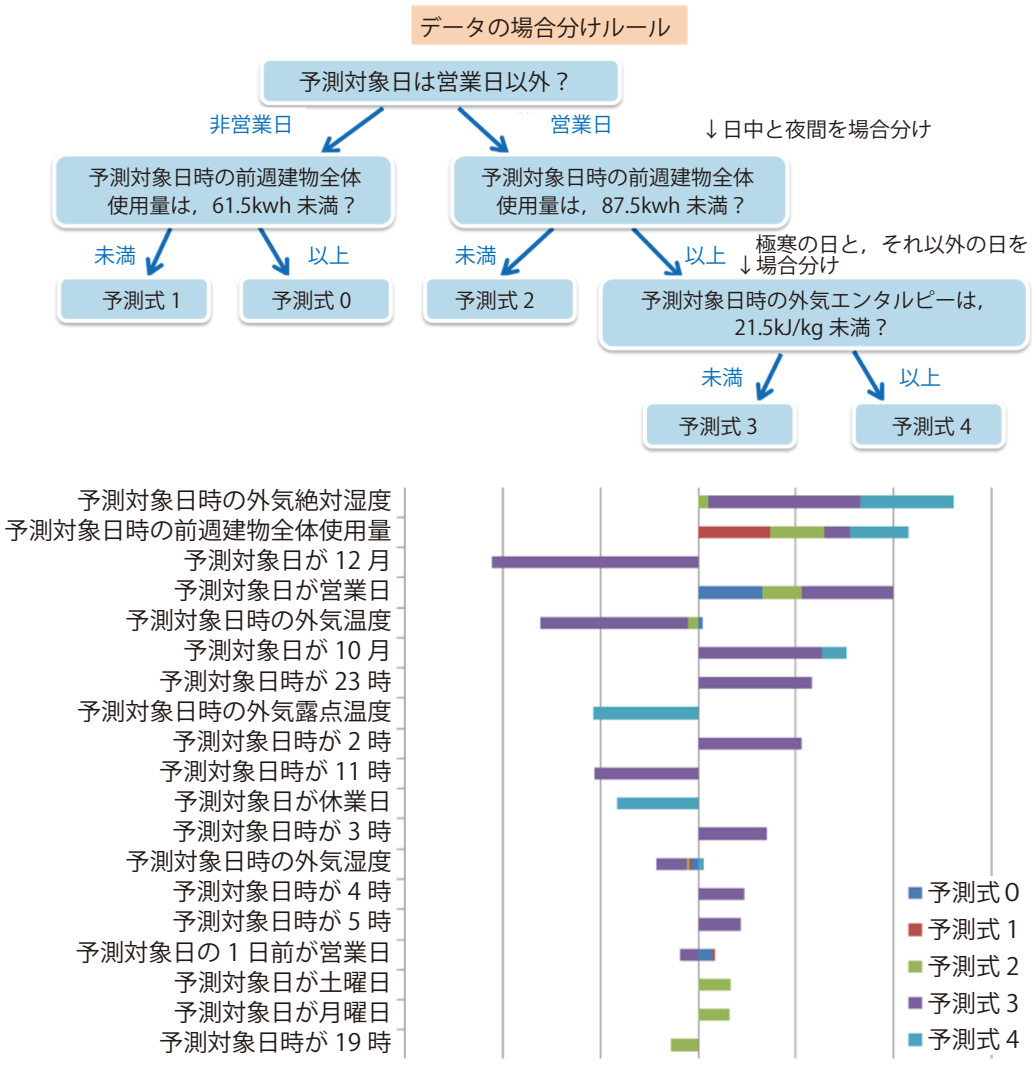

図9　異種混合学習で得られた予測モデル
上図：データの場合分けルール
下図：予測式（横軸は重み係数）

Squared Error）などで計測することができる．これらの指標は定量化が簡単で汎用的に利用可能なため，ビジネス上の効果（たとえばReturn on Investment（ROI）や制約充足）に関する議論が十分になされないままに，精度を追求した結果，分析プロセスとしての性能は良いが，実際の運用には至らないというケースが多くあった．分析プロセスの実運用上の評価指標は，前述の誤差指標以外のものも設定されるべきであり，理想的には図3の7.1（分析計画合意）で事前に顧客と合意形成されるべきである．筆者らは，このプロセスを標準化することで，早い段階から分析の価値を意思決定者（多くは予算執行権限者）へ伝え，スムーズに分析プロセスを運用へ導入することを目指している．

3つ目の課題は，分析プロセスから出力される予測モデルの運用（図2右）である．予測分析の普及に伴って，分析プロセスから出力される分析結果から得られる知見を業務に活かすだけではなく，予測モデルそのものを業務システムに組込む事例が増加している．予測モデルを定常的に運用すると，時間とともに予測精度が悪化するため，モデルをメンテナンスすることがデータサイエンティストの重要な役割となる（たとえば，定期的に再学習する）．しかし，価格予測システムでモデルを再学習すると，モデルの更新前後で価格が変わってしまうといった予測の安定性の問題が起こる．筆者らは，予測モデルの経年劣化を抑えながら，予測の安定性といった運用上の課題を解決し，予測モデルを自動的に運用する方法論について検討を進めている（初期検討の結果に関しては，参考文献[13]を参照）．

5. 説明変数設計自動化への挑戦

異種混合学習によって，データサイエンティストによる予測モデル設計の作成を自動化あるいは大幅に効率化することができるようになった．しかし，このためには，データが正しく処理されて，意味のある情報となって入力される必要がある．すなわち，図1で示したKDD Processにおける前処理工程が重要である．

データの前処理工程には，ETL（Extraction，Transformation，Loading），データクレンジング（異常値や欠損値の除去）などがあり，近年では商用のパッケージソフトウェアをはじめとしたツールが普及し，これらを標準ツールとして整備することで，比較的簡単に処理することができる（図3の7.3や7.4の工程で処理される）．一方で，前節で述べたように説明変数の候補集合を作成する説明変数設計は，依然としてデータや対象ドメインに対するノウハウが必要となる時間のかかる作業である．本章では，筆者らの説明変数設計自動化への取り組みを紹介する．

機械学習分野では古くは主成分分析から，最近では深層学習まで，特徴空間の設計を技術的に解決する手法が精力的に研究されている．一方で，これらの機械学習アルゴリズムによって生成される説明変数（特徴量や属性ともよばれる）は，機械学習や数理統計の専門知識のない一般のユーザに理解することが難しく，4.1節で説明した精度と分かりやすさの両立という観点では問題がある．そこで，データ分析の現場では，業務に関する知識や経験をベースに予測に効果がありそうな説明変数を手作業で設計するということがよく行われている．たとえば，ビールの売上げ予測では，気温が売上と相関を持つことが知られているが，**図10**に示されるように，単に気温データをそのまま使うのではなく，特定の区間を引き伸ばす窓関数を適用することで精度が向上する．

筆者らは，人間に理解可能な説明変数を自動的に設計するための技術開発と，実際のデータ分析プロジェクトへの導入を進めている．我々は，まず説明変数設計の自動化に向けた第一歩として，業務でよく利用される説明変数を標準ツールによって作成できるよう整備した．これは，たとえば移動平均や対数変換といった時系列分析でよく使われるものや，フーリエ変換など周波数分析で利用されるものなど，多数の分析案件で得られたノウハウを基にして，さまざまなドメインの分析に対応可能な最低限のセットとなっている．次に，これらの説明変数を組み合わせることで，予測精度の改善が見込まれる新しい説明変数をデータから自動

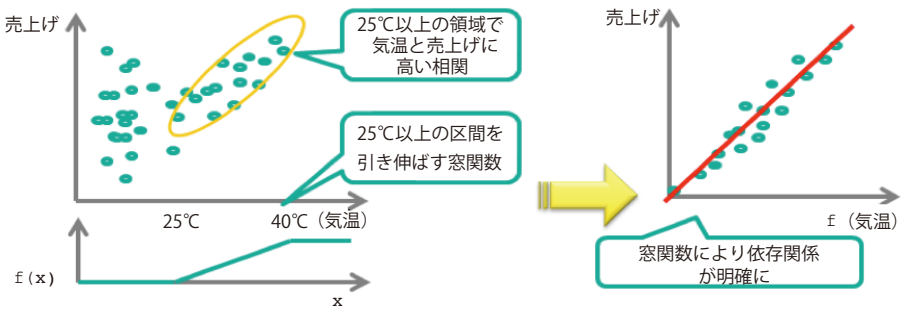

図 10　特徴量設計の例

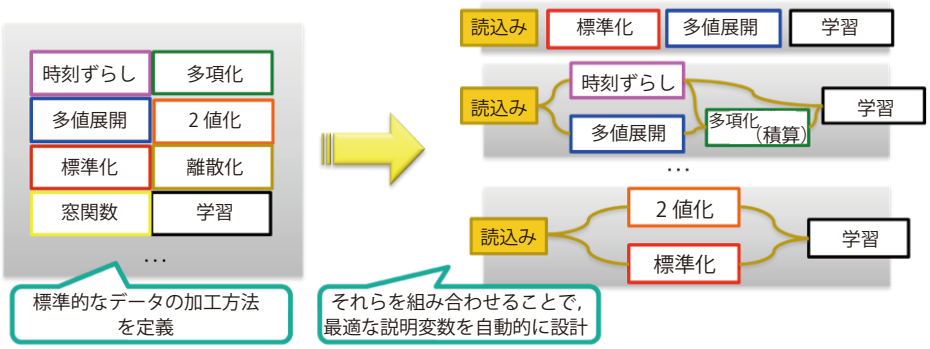

図 11　説明変数の自動設計の模式図

的に生成する技術を開発した（図 11）．説明変数の設計の品質は，データサイエンティスト個人の職人的なスキルに依存していたが，標準整備されたツールと，説明変数組合せを自動化する技術によって，質の高い説明変数を誰もが素早く作成できるようになる．

我々は，電力需要予測☆1 を例にとり，前述の自動設計方式の有効性検証実験を行った．3つのビルの翌日の電力需要を 1 時間単位で予測するという課題について，自動設計方式の有無での精度や工数の差を検証した．精度の比較結果を図 12 に示す．サイエンティストが人手で設計した説明変数を入れた結果と同等以上の精度を自動的に達成し，自動設計を利用した場合，人手による設計の 1/3 程度に工数を削減可能なことが分かった．

6. おわりに

ビッグデータ分析への期待と重要性が高まるなか，データサイエンティストは，さまざまなシーンでデータを通じた課題発見，解決，運用の分析プロセスを素早く行う必要がある．そのためには，データサイエンティスト個人の能力ではなく，一定のスキルを持ったエンジニアであれば質の高い分析プロセスを実行できるさまざまな仕組みが必要である．本稿では，分析プロセスを標準業務フローとして定義する，分析プロセス標準化への取り組みと，分析プロセスの中で個人のスキルへの依存が顕著である説明変数と予測モデルの設計を，技術的に自動化する取り組みを紹介した．本稿が読者の現場のデータサイエンスプラクティス改善の一助となれば幸いである．

☆1　4.3 節とは異なるデータ

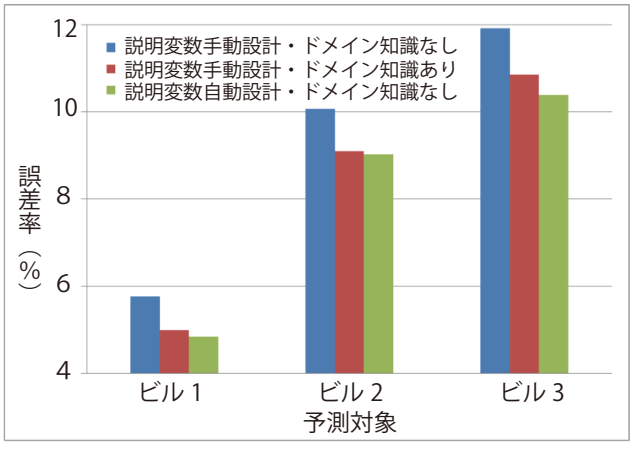

図 12　説明変数自動設計の効果

参考文献

1) Vesset, D., Eastwood, M., Zaidi, A. and Dialani, M.: Worldwide Business Analytics Technology and Services 2013-2017 Forecast, IDC Co. (2013).
2) 平成 24 年版総務省情報通信白書：スマート革命が促す ICT 産業・社会の変革，http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/pdf/n2010000.pdf
3) Fayyad, U., Piatetsky-shapiro, G. and Smyth, P.: From Data Mining to Knowledge Discovery in Databases, AI Magagine, Vol.17, pp.37-54 (1996).
4) Fairness, Acountability, and Transparency in Machine Learning, http://www.fatml.org/index.html
5) Rudin, C.: Algorithm for Interpretable Machine Learning, Invited Talk in 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2014). http://videolectures.net/kdd2014_rudin_machine_learning/
6) Big Data: Seizing Opportunities, Preserving Values, Exective Office of the President (White House) (2014).
7) Fujimaki, R. and Morinaga, S.: Factorized Asymptotic Bayesian Inference for Mixture Modeling, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS) (2012).
8) Fujimaki, R. and Hayashi, K.: Factorized Asymptotic Bayesian Hidden Markov Model. Proceedings of the 25th International Conference on Machine Learning (ICML) (2012).
9) Hayashi, K. and Fujimaki, R.: Factorized Asymptotic Bayesian Inference for Latent Feature Models, 27th Annual Conference on Neural Information Processing Systems (NIPS) (2013).
10) Eto, R., Fujimaki, R., Morinaga, S. and Tamano, H.: Fully-Automatic Bayesian Piece-wise Sparse Linear Models, Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) (2014).
11) Liu, J., Fujimaki, R. and Ye, J.: Forward-Backward Greedy Algorithms for General Convex Smooth Functions over a Cardinality Constraint, Proceedings of the 27th International Conference on Machine Learning (ICML) (2014).
12) 藤巻遼平，森永　聡：ビッグデータ時代の最先端データマイニング，NEC 技報，Vol.65, No.2 (2012).
13) 谷本　啓，本橋洋介：モデルのライフサイクルを考慮した大量予測モデル管理手法の検討，第 29 回人工知能学会全国大会論文集 (2015).

藤巻 遼平（非会員）rfujimaki@nec-labs.com
2006 年東京大学大学院工学系研究科航空宇宙工学専攻修士課程修了．NEC 情報・ナレッジ研究所主席研究員，北米分室（ビッグデータ分析）リーダ．機械学習・データマイニング原理の研究開発とビッグデータ分析の事業化・海外展開に従事．工学博士．

本橋 洋介（正会員）y-motohashi@bk.jp.nec.com
東京大学大学院工学系研究科産業機械工学専攻修了．2006 年日本電気（株）入社．NEC 中央研究所で機械学習・データマイニング・ナレッジマネジメント・コラボレーションソフトウェアの研究開発に従事．近年は，機械学習の研究開発成果を活用した分析案件を推進．2014 年よりビッグデータ戦略本部兼務．

採録決定：2015 年 6 月 9 日
編集担当：住田一男（(株）東芝）