特集論文

AIとビッグデータ時代を担うデータサイエンティストの育成とデータ分析の実践

中村伊知郎* 小林 敦**
吉原晴香** 松岡誠二**
白浜広彬**

Development of Data Scientists Responsible for Era of AI and Big Data, and Practices of Data Analysis

Ichiro Nakamura, Haruka Yoshihara, Hiroaki Shirahama, Atsushi Kobayashi, Seiji Matsuoka

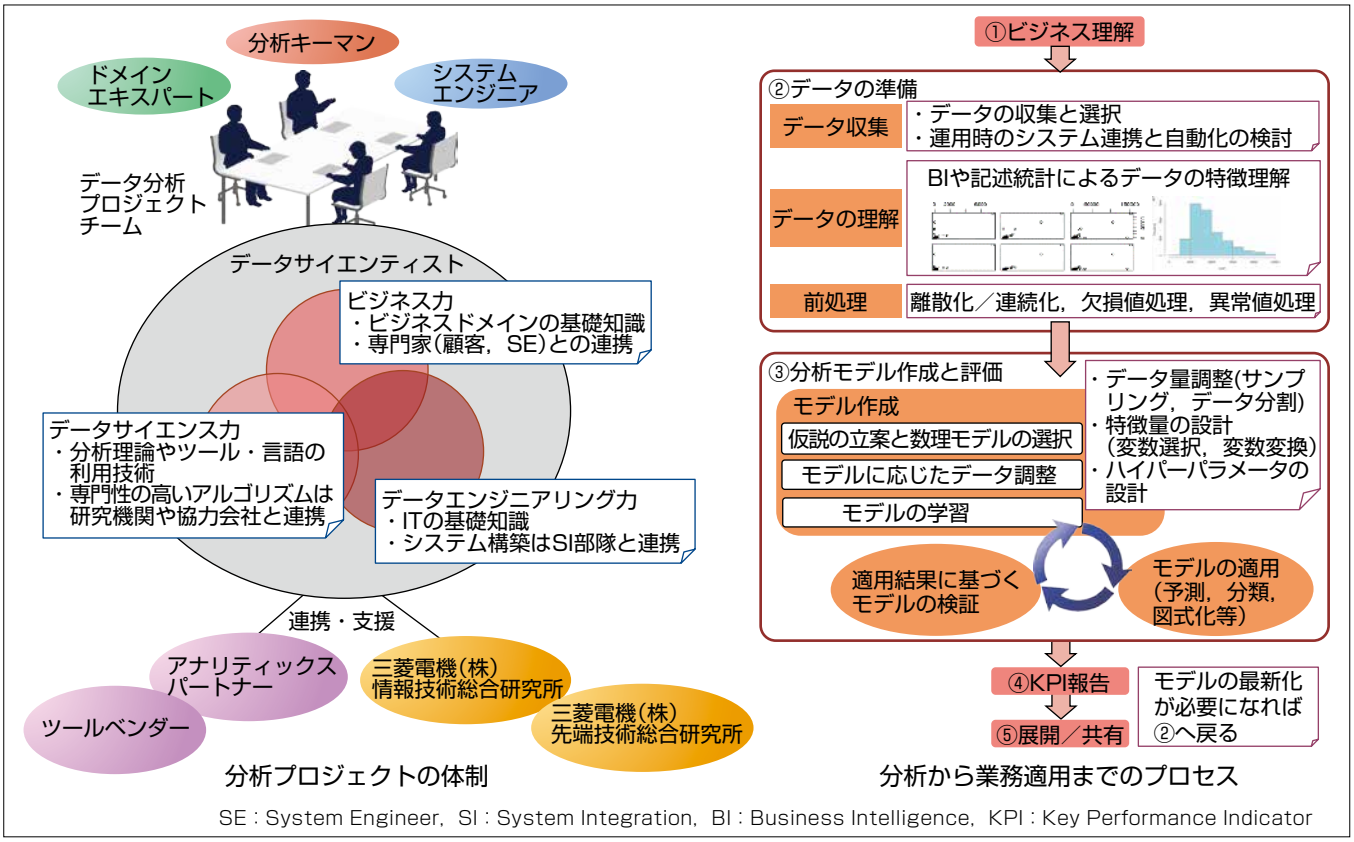
要 旨

近年，ITの処理能力の向上に伴い，蓄積されたビッグデータから新たな価値を見い出すことが，企業にとって競争力の源泉になると言われている。三菱電機インフォメーションシステムズ㈱(MDIS)では，データから価値を見い出し，新たなビジネスを顧客と創出することを目的に，データサイエンティストの育成を2013年から開始した。

育成方針としては，分析ツールのプロフェッショナルとしてではなく，統計学や機械学習理論の基礎を身につけた上で，データの背後にある意味・構造の解明や，新しい手法への対応も可能な人材育成とした。育成方法は，理論面の教育，言語・ツールの教育，OJT(On-the-Job Training)の三段階で行っている。

データ分析の実践としては，ウェアラブルデバイスを用いた疲労要因分析，設備故障の遠隔復旧実施可否の推定などが挙げられる。そのほかにも様々な業種で，需要予測，品質管理，IoT(Internet of Things)データに基づく予測・診断等があり，その内容もディープラーニングを中核としたAI(Artificial Intelligence)や，大量の時系列データ処理等へ変化してきているが，これまで進めてきた育成方針が実を結び，これらの変化にも柔軟に対応し，成果を出している。

システムインテグレータであるMDISには，顧客が持つデータから新たな顧客価値を創造することが求められている。常にデータと向き合い，そこから真理を導き出し，経営や業務を進化させ続ける企業文化を，顧客の中に定着させていく。



SE : System Engineer, SI : System Integration, BI : Business Intelligence, KPI : Key Performance Indicator

データ分析のプロジェクト体制と分析から業務適用までのプロセス

一般社団法人 データサイエンティスト協会の定義に従うと，データサイエンティストは，“ビジネス力” “データサイエンス力” “データエンジニアリング力”の三つの力を必要とする(1)。データ分析プロジェクトの進行には，ドメインエキスパート，システムエンジニアや研究機関との連携が不可欠となる。分析の実施は，データの準備段階で多くの時間を必要とし，その後も有用な知見を得られるまで試行錯誤を続けるプロセスになる。

*三菱電機インフォメーションシステムズ㈱(博士(工学)) **同社

1. まえがき

経営学者の伊丹敬之[2]は情報を“見えざる資産”と称し，情報が経営にもたらす価値を説いてきた。近年はさらにITの処理能力の向上によって，データから新たな価値を見い出すことが企業の競争力の源泉となり，データを握るものがビジネスを握ると言われている。以前と異なるのは，データが多量(ビッグデータ)であること，そして分析手法の高度化・複雑化である。MDISは，オープンソースソフトウェアHadoop(注1)やSpark等を応用したビッグデータ処理の仕組みを顧客に納入する従来型のシステムインテグレーション事業に加えて，蓄積されたデータから価値を見い出し，新たなビジネスを顧客とともに創出する役割としてデータサイエンティストを擁し，顧客にデータ価値を提供するという従来とは異なる事業の在り方へ挑戦している。

本稿では，6年間にわたるデータサイエンティストの育成の経過とMDISが実践した事例の中から，顧客の業務の現場でのデータ分析の取組み事例を述べる。

(注1) Hadoopは，The Apache Software Foundationの登録商標である。

2. データサイエンティストの育成

MDISでは，2013年からデータサイエンティストの育成を開始した[3](**図1**)。ITシステムの歴史や昨今の進化を考えれば，データ分析に対する計算機支援も急速な高度化や自動化が予想されたため，そうした流れの中でも長期的に活躍し続けられる人材の育成を目指している。選抜された若手エンジニア数名を対象に育成を始め，その後は，大学でデータサイエンスを学んだ新人を登用し，育成している。

2.1 育成方針

データ分析の多くは専用のツールを使うことで，その理論を十分理解せずにブラックボックスとして扱っても，何らかの結果を得ることができる。しかし，その結果が顧客の期待とは異なるものになった場合は，利用データや分析手法に踏み込んだ調査を行って，その原因と改善可能性を説明できる必要がある。また，単純な問題であれば機械学習を汎用的に利用できるが，各ビジネスドメインには背景にある数理的構造の解明にまで踏み込んだ独自の分析手法が発展していることが多く，これらを使いこなすことも必要になる。こうした背景から，分析ツールのプロフェッショナルを育成するのではなく，統計学や機械学習理論の基礎を身につけた上で，データの背後にある意味・構造の解明や，新しい手法への対応も可能な人材を育成する方針としている。

2.2 育成の三つの柱

MDISのデータサイエンティスト育成は，**図1**に示すように，理論面の教育に加えて，言語やツールの教育，及びOJTの三つの柱からなる。また，参加プロジェクトで,“ビジネス力”や“データエンジニアリング力”の基礎スキルを，ドメインエキスパートやシステムエンジニアから指導を受けている。

2.2.1 統計学や機械学習理論の教育

数理統計や統計的検定手法等の理論面での教育は，社内輪講を中心に，社外セミナ等も利用し，一般財団法人 統計質保証推進協会の統計検定2級／準1級の取得を促進している。また，機械学習やディープラーニングに関しては，2018年から開始された一般社団法人 日本ディープラーニング協会の認定プログラムも活用し，E(エンジニア)資格の取得を促進している。

2.2.2 分析言語やツールの教育

計算機上での分析実施スキルに関しては，統計用プログラミング言語を使った分析から習得する方針としている。これは，最初からGUI(Graphical User Interface)型

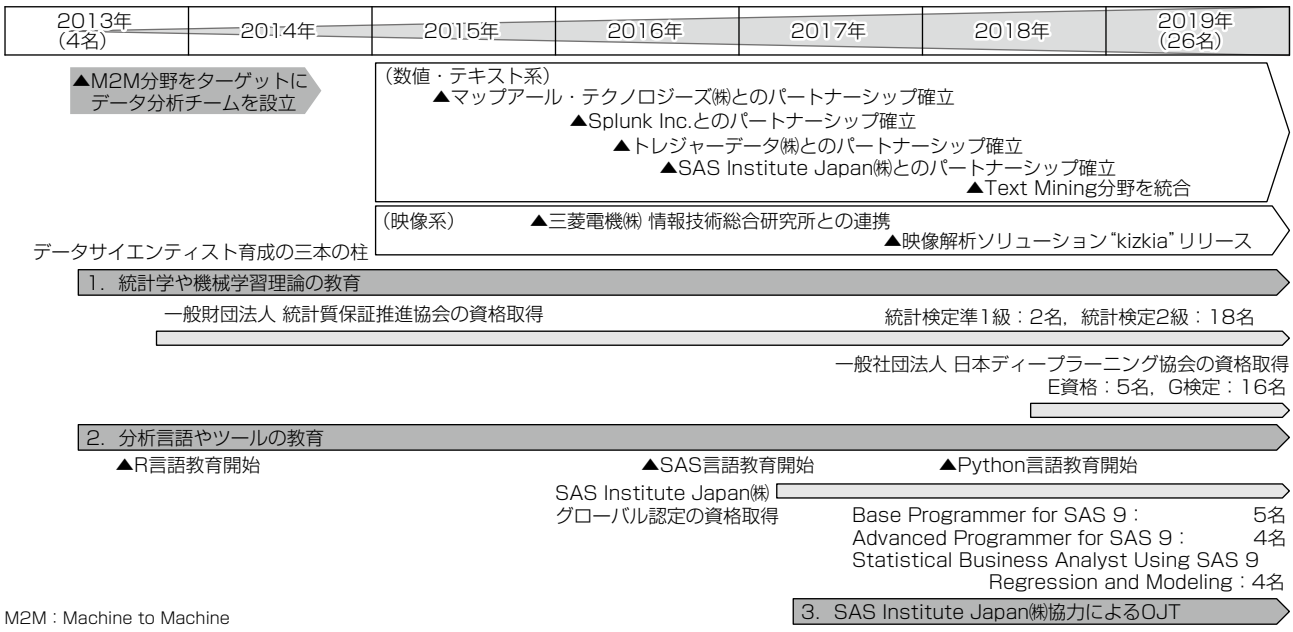


図1．MDISのデータサイエンティスト育成の経緯

分析ツールに慣れてしまうと，理論との対応付けを考えるトレーニングができないためである。まず，R言語とPython言語(注2)を中心にトレーニングを進め，データベースやWebの開発も経験させ，分析プロセスと理論との対応付けができるようになった後に，GUI型分析ツールの習得を行うことで，背景となる理論の理解を深めるように進めている。

データ分析プロセスは試行錯誤が多く，言語主体での分析ではコーディングが煩雑で生産性が低くなるため，GUI型分析ツールの利用による効率化が重要となる。また，運用時には性能や信頼性に加えてサポートの継続性が要求されるため，商用製品の利用が望ましい。一方で，GUI型分析ツールは典型的な用途に限定して実装されているため，統計用言語による機能開発と柔軟に組み合わせて使う必要がある。こうした背景から，世界的評価が高く，ビジネスドメインごとにGUI型分析ツールが用意され，さらにプログラミング言語による機能拡張も可能なSAS Institute Japan㈱(SAS社)の製品を採用し，そのグローバル資格の取得を促進している。

(注２)　Pythonは，Python Software Foundationの登録商標である。

2.2.3　SAS社の協力によるOJT

机上で一通りの教育を終えた後は，身に着けた理論や分析スキルの幅を広げ，また当人が自信を持って実践につなげていけるように，データ分析を専門としている企業でのOJTを行っている。SAS社の協力も得て，幅広い業種と分析内容でのデータ分析を経験しており，受注予測，消費者生活行動予測や論文等で発表された新しい分析技術の実装等を行っている。

2.3　社内コミュニティの設立

その後，MDIS内でデータ分析に取り組む事業部門が複数に広がってきたため，共通して育成に貢献できる社内コミュニティを設立した。理論や分析実施スキルの輪講を中心に，個別案件で採用する分析手法の議論や勉強会，リソースの相互支援や関連製品の紹介等，活動は多岐にわたっており，成長しあえる関係を築いている。

3.　データ分析の実践

MDISのデータサイエンティストは，製造，設備管理，流通・小売，通信等の様々な業種で，需要やイベントの予測，品質管理，IoTデータに基づく予測・診断等の多様な分析を実施してきた(**表１**)。ここでは，保守サービスでの二つの分析事例について述べる。

3.1　ウェアラブルデバイスを用いた疲労要因分析

現場作業に対する安全管理の一環としてウェアラブルデバイスの活用を検討した疲労要因分析の事例について述べる。

3.1.1　案 件 概 要

図２に示すように，ウェアラブルデバイスから取得した“バイタルデータ”に加えて，作業報告から得られる“業務データ”と“疲労度のアンケート”結果，及び身体情報として“保守員プロフィール”のデータを入手し，それらの関連性を分析して，保守員が疲労を感じる要因を特定できるかを検証した。

疲労度アンケートで“疲労あり”の回答が高かった作業時間帯で，バイタルデータに含まれる身体負荷指数(年齢及び安静時／現在の心拍数から推定した身体に掛かる二つのデータを，保守員の疲労度を表すデータとして扱う負荷の度合い)も高い値を示すことが確認できたため，この二つのデータを，保守員の疲労度を表すデータとして扱うことにした。その他のデータから，疲労を感じる要因の候補となる項目(以下“疲労要因候補項目”という。)を抽出した。

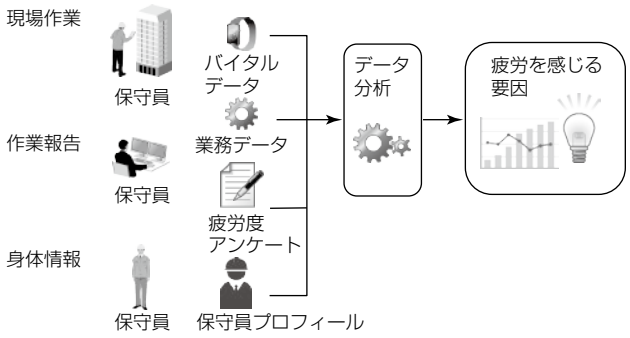


図２．疲労要因の分析

表１．MDISのデータ分析事例(抜粋)

業種	分析内容	言語／ツール	解析手法
保守サービス	ウェアラブルデバイスを用いた ・作業者支援 ・疲労要因分析	Power BI(注3), SAS Enterprise Guide(注4)(EG)	異なる尺度変数間の相関分析，決定木
保守サービス	・設備の故障解析 ・設備故障の遠隔復旧実施可否の推定	R，Python，Keras＋TensorFlow(注5)	ロジスティック回帰，ランダムフォレスト， 勾配ブースティング，深層学習(CNN)
製造	顧客問合せデータの可視化	Tableau(注6), SAS Visual Analytics(VA)	部門間の流量分析(サンキー・ダイアグラム)， Webページ間の参照分析(ネットワーク分析)
製造	生産ラインのオペレータ支援と 自動化検討	SAS Enterprise Miner(EM), SAS Event Stream Processing(ESP)	決定木，重回帰分析，時系列解析
製造	少量生産品の需要・出荷傾向分析	SAS EG，SAS/QC(注4)	管理図，非線形回帰(ポアソン分布)
製造	製品需要予測	SAS EG，SAS Forecast Server	時系列解析，非線形回帰(ゴンペルツ曲線)
IT	ITインフラ提供サービスの情報提供	SAS EG，SAS VA	可視化，問合せ間の流量分析
サービス事業者	消費者生活行動予測	SAS EG，SAS EM	ロジスティック回帰，ランダムフォレスト， 勾配ブースティング
化学	製品開発での需要予測	SAS EG，SAS Forecast Server	時系列解析，非線形最適化(コスト⇔売上)， グラフィカルラッソ(変化検知)

(注3)　Power BI は，Microsoft Corp.の登録商標である。
(注4)　Enterprise GuideとSAS/QCは，SAS Institute Inc.の登録商標である。
(注5)　TensorFlowは，Google Inc.の登録商標である。
(注6)　Tableauは，Tableau Software Inc.の登録商標である。

3.1.2 疲労要因項目の重要性の分析

最初に，疲労要因候補項目の中から，疲労との関連性が高い項目の絞り込みを行った。**図3**に示すように，縦軸に示した疲労要因候補項目ごとに，横軸に示した“疲労あり”回答と身体負荷指数との関連性を調査した。カテゴリカルデータと数値データが混在するため，データ型の組合せに応じて相関係数，相関比，クラメールの連関係数の三種類の相関係数を算出した。各係数の強弱の指標を使って，“疲労あり”回答と身体負荷指数が，各疲労要因候補項目に対して，同じ関連性の傾向を示すかを評価した。**図3**に示すように，“疲労あり”回答と身体負荷指数の両方と関連性が高い項目もあれば，“疲労あり”回答との関連性だけが高い項目もあった。最終的には，担当部門の主観との一致性も考慮して，疲労要因候補項目を絞り込んだ。

3.1.3 疲労要因項目の優先度や境界値の分析

次に，“疲労あり”回答との高い関連性を示した疲労要因候補項目を中心に，疲労度の判断に使う際の優先順位と，判断の基準になる境界値を調査した。決定木系のアルゴリズムを使って分析した結果，作業負荷が比較的高いと保守員が感じる機種や作業内容等を特定できる結果が得られ，その特定結果が担当部門の主観と一致することが分かった。

以上のように，疲労に関する二つの計測データ（“疲労あり”回答，身体負荷指数）と各疲労要因候補項目の関連性から，疲労度判断に有用な疲労要因項目を割り出し，最終的に疲労度を判断する際の疲労要因項目の利用順位と境界値を導出した。各段階で，分析結果が担当部門の主観と一致することを逐次確認しながら，分析の深堀を進めた。

3.2 設備故障の遠隔復旧実施可否の推定

設備故障の遠隔復旧実施可否に関して，データ分析による推定を検討した事例について述べる。

3.2.1 案 件 概 要

図4に示すように，異常発生時の設備の各種センサの状態を記録した“センサデータ”に加えて，問合せ内容や遠隔監視装置からの情報を記録した“故障情報”，設備の情報等を記録した“各種台帳”，及び正解データとして“設備故障の遠隔復旧実施可否の判断例”の正解ラベルを入手し，データ分析によって設備故障の遠隔復旧実施可否の判断を再現できるかを次のように検証した。

3.2.2 設備故障の遠隔復旧実施可否の推定

図5に示すように，この分析に用いるデータは，状態の時間的な変化を記録した時系列データもあれば，時間的な変化のないデータもあり，また時系列データの長さも一定ではないという特徴を持つ。このような可変長の値が複数含まれるデータは，従来のどの分析手法にも適さない。そこで，この分析ではセンサデータの持つ情報を重視する方針をとり，補完等によって各特徴量の時間軸方向の長さをそろえる手法を採用した。

図6にディープラーニングを用いた場合の推定処理の流れを示す。この分析では各特徴量で時間軸方向のデータの補完を行っているため，時間軸上の位置に対するロバスト性を高めることで推定精度の向上を図った。具体的には，プーリング処理や畳込み処理が可能なCNN（Convolutional Neural Network）モデルを採用することによって，先行して実施したロジスティック回帰による推定精度を大きく改善できた。

CNNによる推定精度は，正確度（Accuracy）の面では80％を超えているが，F値（F-measure）の面では，まだ課題が多い。これは，①対象とした設備は故障が少ないため，予防点検作業や訓練等による停止を除いてしまうと，

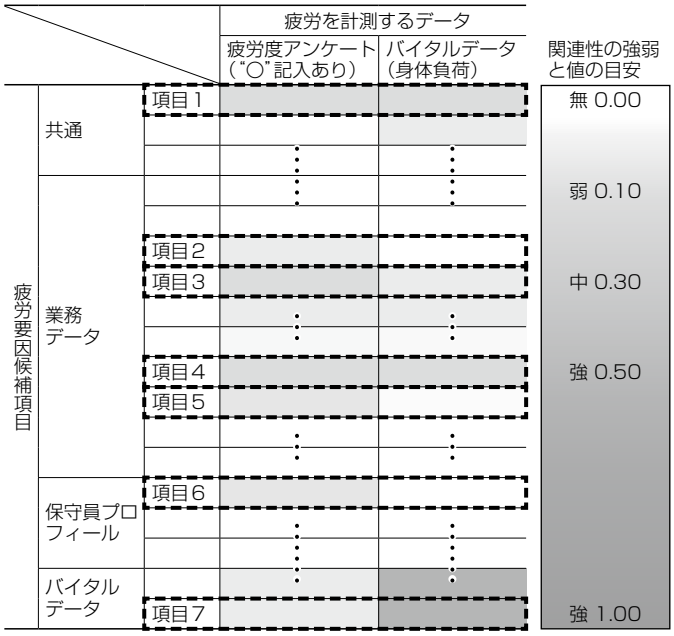


図3．疲労要因候補項目ごとの疲労計測データとの関連性

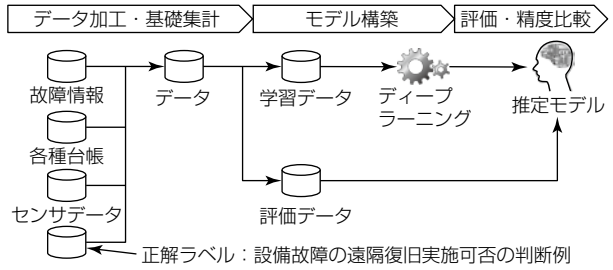


図4．設備故障の遠隔復旧実施可否の推定モデル構築

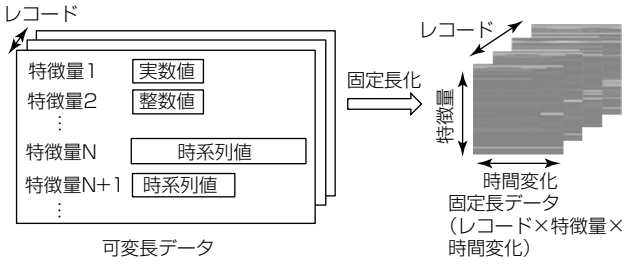


図5．特徴量の固定長化

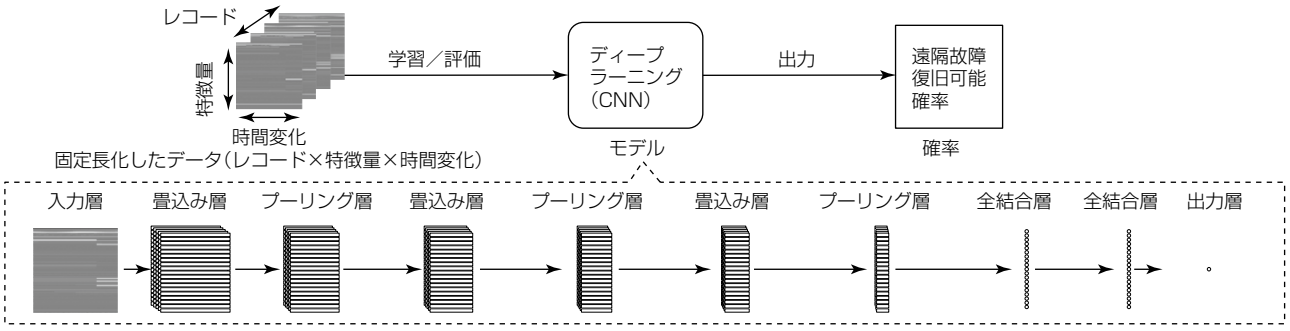


図6．ディープラーニングを用いた推定処理の流れ

遠隔故障復旧に該当する学習データが少ない点や，②“設備故障の遠隔復旧実施可否の判断例”の正解ラベルには保守員による判断が含まれるために揺らぎが少なくない点等の原因による。①は今後のデータ収集によって解決し，②は正解ラベルの精度向上を進める方針とした。

3.2.3　停止要因によるデータのセグメンテーション

停止要因が同一のデータ同士は，その特徴量も類似している可能性が高く，正解ラベルも同一である可能性が高い。そこで，データを設備の停止要因ごとにセグメンテーションし，その結果を，正解ラベルの付与条件の精査に利用することで，正解ラベルの揺らぎを減らそうとしている。セグメンテーションの学習データには，作成したCNNモデルの中間層の重みベクトルを採用した。

以上のように，設備のセンサデータと，故障情報，各種台帳のデータを使って，設備故障の遠隔復旧実施可否の推定モデルを構築した。今後は，推定精度の向上を追求して，設備故障に関する高度な遠隔復旧支援を可能にすることを目指している。

4. データ分析分野での変化への対応

データサイエンティストの育成開始から6年がたった現在のデータ分析分野での変化として，ディープラーニングを中核にしたAIの普及や，IoTや機器が生み出す時系列データの増加等が挙げられる。

ディープラーニングや高度な機械学習等の新しい手法は，予測や分類の精度向上が期待できる一方で，結果説明性に乏しいという問題を残している。また，データ自体に不備や冗長性等の問題を抱えていることも多く，これら新しい手法でも十分な分析精度を確保できないことがある。MDISのデータサイエンティストは，これら新しい手法と従来手法を組み合わせた分析を行うことによって，結果に至った理由を顧客に分かりやすく提示することに取り組んできた。また，不備や冗長性のあるデータに対しても，データの分布，交互作用や因果関係にまで踏み込んで，特徴量の設計や変換を実施することによって，分析精度の向上に取り組んできた。

一方，IoTや機器が出力する時系列データは，周期性や規則性を期待できないことが多く，従来の時系列解析の手法だけでは処理できない場合がある。MDISのデータサイエンティストは，前提にできる確率分布の利用，ベイズ的な手法の適用や，時間軸に沿った相関関係や因果関係の変化検知等の，新旧の手法を臨機応変に組み合わせることによって，こうした時系列データの分析に取り組んでおり，幾つかの案件で成果を出してきた。

このように，データ分析分野での最近の変化にも柔軟な対応ができており，**2. 1節**で述べた育成方針が実を結んできていると言える。一方で，データ分析での新たな課題や新しい手法が絶え間なく生まれており，今後も育成プログラムを随時見直していく必要性を認識している。

5. むすび

現在，ビッグデータとAIの領域で経営的成果を挙げている企業の多くは，自社内にデータの発生源があり，それを活用すること自体を自社の経営戦略に組み込んで，自社のデータサイエンティストが分析業務を行うという自己完結型である。一方でMDISのようなシステムインテグレータは，顧客が持つデータの中から新たな顧客価値を創造することが求められ，さらに顧客の中にデータ活用の仕組みや体制を構築し，データ分析業務を定着させてゆくことが求められる。大切なのは，常にデータと向き合い，データの中から真理を導き出し，経営や業務を絶え間なく進化させ続ける企業文化を顧客の中に作り上げることである。MDISは，データサイエンティストの活動をもってそれを支援していく。

参考文献

(1) データサイエンティスト協会：スキルチェックリスト ver2.00 (2017)
https://www.slideshare.net/DataScientist_JP/2017-81179087
(2) 伊丹敬之：経営戦略の論理，日本経済新聞出版社 (1980)
(3) 尾崎　隆，ほか：データサイエンティストとM2M（ビジネス・技術・育成），オペレーションズ・リサーチ，**59**，No.9，543～548 (2014)