いまさら聞けない機械学習入門（後編）：

機械学習はどうやって使うのか――意外と地道な積み重ね

http://monoist.atmarkit.co.jp/mn/articles/1805/22/news010.html

前編では、AI（人工知能）と機械学習、ディープラーニングといった用語の説明から、AIを実現する技術の1つである機械学習が製造業を中心とした産業界にも徐々に使われ始めている話をした。後編では、機械学習を使ったデータ分析と予測モデル作成について説明する。

2018年05月22日 10時00分 更新

[西啓（PTCジャパン株式会社）, MONOist]

1. IoTと機械学習

前編では、AI（人工知能）と機械学習、ディープラーニングといった用語の説明から、AIを実現する技術の1つである機械学習が、製造業を中心とした産業界にも徐々に使われ始めている話をした。背景に少し触れると、産業界に訪れたIoT（モノのインターネット）の普及により産業機器のデジタルデータ化が進んだことで、機械学習に必要となる大量のデジタルデータの提供が可能になったことがあげられるだろう。機械学習で作成した予測モデルをIoTアプリケーションに結合して、リアルタイムでの予測機能を実現する事例も出てきている。

AIのイメージが肥大化しているせいか、データさえ用意すれば機械学習ツールが自動で何でもやって予測してくれるイメージを持たれているかもしれない。しかし機械学習という言葉とは裏腹に、そのプロセスは意外と地道な手作業と試行錯誤の積み重ねである。そこで今回の後編では、機械学習を使ったデータ分析と予測モデル作成について説明したい。なお機械学習ツールとしてはPTCの「ThingWorx Analytics」を使用している。

2. データの収集と整理

機械学習をやってみようと言っても、とにかくデジタルデータがないと始まらない。機器のデータを取得するといってもどんな種類のセンサーをどこに取り付けるのか、収集の頻度をどうするのか、収集する期間はどのくらいにするのか、といったパラメーターは作業を知る人間が決めるしかない。ツールベンダーによっては特定の作業の知見を持っていることもあるが、ほとんどの場合は現場によって異なるので試行錯誤をするしかない。

また作業者による記録、顧客からの不具合の通知内容など、別々のソースから上がってくるデータをどのようにまとめるのかも、自動ではできずに人間が決める必要がある。秒単位で上がってくるセンサーデータと、時間単位や日単位になる手入力データを合わせるとしたら、時間軸をどうするか。また、記入ミスを訂正したり、データに欠落があった場合にはその記録をどう

取り扱うのか（0を入れるか、平均値を埋めるのか、削除するのか）、重複するデータの削除といった地道な作業が必要となる。

そして何よりも重要なのは、ゴール変数の設定である。機械学習のゴールは連続値（いわゆる数値）もしくは文字列により、明確に定義する必要がある。何を知りたいのか、また集めたデータから何を読み取れそうか、といったところを考慮して人間が決めなくてはならない。

例えば、ある機器が設置現場からメンテナンスのために戻して工場で実施した検査結果のデータを用いて、90日以内に起きる機器の故障の予兆を診断するシナリオを考えてみよう（図1）。ここでのゴール変数は1（故障する）か0（故障しない）と定義している。しかし、単純に検査結果がNGのときにゴール変数を1としてしまうと、予兆は予測対象とならず、検査結果がNGとなる予測になってしまう。90日以内に起きる予兆を知りたいのであれば、同一シリアル番号の機器における90日以内の各項目の値にこそ予兆が隠れているだろう。従って、図1の下の表にあるように、2017年12月10日の検査結果OKのレコードのゴール変数が1（故障する）となる。

図1　90日以内に起きる機器の故障の予兆を診断するシナリオ（クリックで拡大）

実はこのデータの準備こそが一番手間が掛かる上に自動化が難しい。しかもデータは重要だ。なにせ、この後の分析／予測では、用意したデータこそが機械学習ツールにとっての世界の全てとなるのだ。

人間であれば長年蓄えてきた経験や知識といった基礎データがあるが、機械学習ツールは入力したデータだけが分析対象となる。しかも、対象となるデータの件数はある程度の量が必要となる。これもケース毎に異なるので、1000〜10万件以上と幅は広いが、いずれにせよ大量のデータを整理する必要がある。最近ではデータの編集作業に特化したデータラングリング（Data Wrangling：直訳すると「データを飼いならす」）ツールが登場しており、編集ルールをプログラムできるようになってきている。ただし、そのルールを決めるのはあくまで人間であり、最初は手探りの作業であることは心に留めておいてほしい。

3. 集めたデータの分析

データを整理し機械学習ツールの形式に成形したところで、ようやくツールの出番となる。まずは集めたデータを分析し、ゴール変数に対してどの入力変数がどのくらい関係しているのかを見てみる。

図2の例では、ThingWorx Analyticsが持つシグナル機能により、エスプレッソマシンのグラインダーの故障に、どの要因が関係してるかをランクが高い順に表示している。要因のトップは1日当たりの平均使用回数だ。これだけを見ると当たり前すぎて何の知見も得られないと思われるが、画面右側には故障率順に使用回数のヒストグラムが表示されており、350.5回を境に故障率がおよそ24%から60%に跳ね上がることが読み取れる。このように具体的な数値として要因を分析できるのがツールの便利なところである。

図2　エスプレッソマシンのグラインダーの故障要因（クリックで拡大）

さらに複数の入力変数の組み合わせによりゴール変数の値が高く（もしくは低く）なるグループを探し出すのが、ThingWorx Analyticsのプロファイル機能だ。図3では、先のエスプレッソマシンのグラインダーの故障に一番関わるグループを示している。画面では1日の平均使用回数が350.5～483回のときにおよそ66%の故障率だが、グラインダーがサテライトタイプだと83%に上がると表示されている。使用回数とこのグラインダーのタイプに何か問題があると推測され、設計や顧客への利用ガイドの検討材料となるだろう。

図3　故障率との関係の強い組み合わせ（クリックで拡大）

これらの分析から、入力変数を見直すこともできるだろう。特に、入力変数の数が多い場合は、ゴール変数と関係の弱い入力変数を、この後おこなう予測モデル作成対象から外すことで予測モデルの作成時間を短縮できる。

4. 予測モデルを作る

データがそろったところでようやく予測モデルの作成に移る。一般的な機械学習ツールでは複数の学習アルゴリズムを用意してあり、対象データの傾向やゴールの種類に応じてユーザーが学習アルゴリズムを選択する。作成したモデルを評価するためには、あらかじめデータの2割程度を検証用にとっておき、残りの8割を学習用として使う。

ThingWorx Analyticsではアルゴリズム選定を簡易にするために、数種類のアルゴリズムそれぞれで予測モデルを作成させて競わせて、それらの内ベストな成績を残したモデルを採用する、または上位数個のモデルを採用してアンサンブルモデルとする機能がサポートされている。

予測モデルを評価するためにはその指標が必要となる。連続値の予測の場合はRMSE（平均二乗誤差の平方根）を用いることが多い。RMSEでは0に近いほど誤差が少ない（図4）。

図4　連続値の予測モデルの性能評価にはRMSEを用いることが多い（クリックで拡大）

分類の場合は別の指標を使う。ここでは二項分類、1か0のどちらかに分類する場合を紹介する。二項分類ではROC（Receiver Operating Characteristic）という指標を用いる。二項分類において、単に予測と実際の結果が一致するだけでは、評価が十分ではない。例えば、機器の故障を予測する場合、いつも故障すると予測していれば実際に故障するときに必ず正解する。しかし、実際には故障しないときにも故障すると予測しているので、オオカミ少年となってしまい誰も予測を信じなくなる。

逆に、いつも故障しないと予測していて実際に故障してしまうと、予測は役に立たなくなる。これら2つの予測の出し方は極端としても、故障の予測的中率を上げつつ、オオカミ少年にならないようにするのが理想的な予測モデルである。そこで、横軸にオオカミ少年となる偽陽性率を、縦軸に正解を当てる真陽性率をとる「ROC（受信者動作特性曲線）曲線」を作成し、どのようなカーブの形を描くのかを観察する。

図5の右側のグラフは青い部分の面積が大きいほど優れた予測モデルである。なお青い部分の面積はAUC（Area Under Curve）と呼ばれ、0から1の値を取る。AUCが1に近いほど予測モデルは優れていることになる。

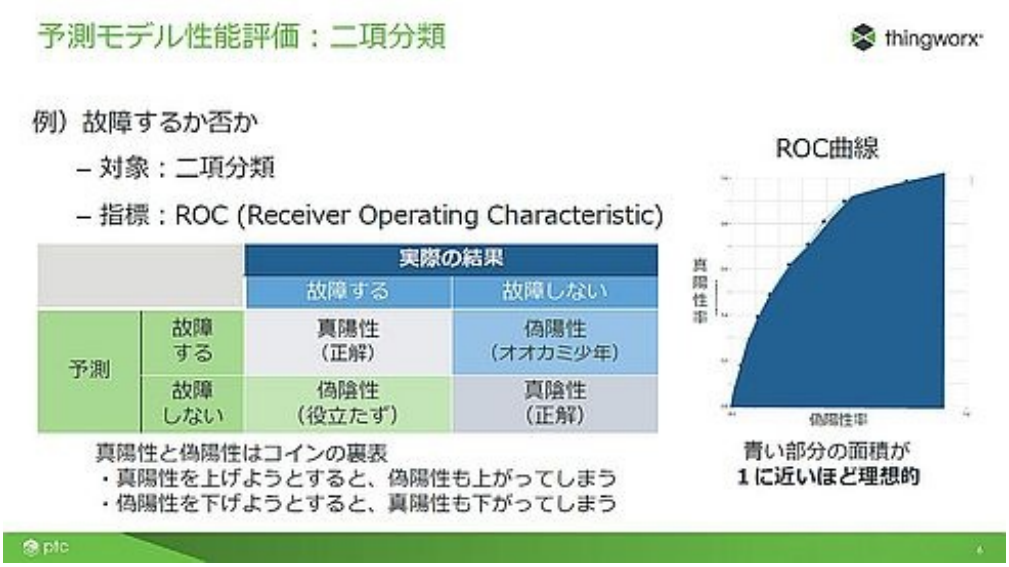

図5　分類の予測モデルの性能評価に用いられる二項分類（クリックで拡大）

5. 作成した予測モデルをどう活用するのか

予測モデルを作成したら実際のデータを入力して予測スコアを出してみる。最近のツールでは、ただ予測スコアを算出するだけではなく、各入力データに対してなぜその予測スコアになったのか説明する機能を持つものが出てきている。

前編では、機械学習について説明よりも予測に重きを置いていると述べたが、ユーザーの要望として全くのブラックボックスを使うことへの不安もあるため、何らかの説明が要望されているのだ。ThingWorx Analyticsでは、入力変数のうち、指定した1〜5種類の重要な変数及びその値を感度分析により提示する機能を備えている（図6）。

図6　予測スコアと、感度分析によって提示された入力変数と値（クリックで拡大）

予測モデルは、作成の際に使用したデータの傾向のままであれば問題無いが、時間の経過や諸条件の変化と共に予測精度が低下していくことが多い。この予測モデルの精度のモニタリングと見直しも人間が担う必要がある。新たなデータを追加して、予測モデルの再学習をする

だけで済むこともあれば、入力変数を追加／削除してモデル作成をやり直すこともある（図7）。

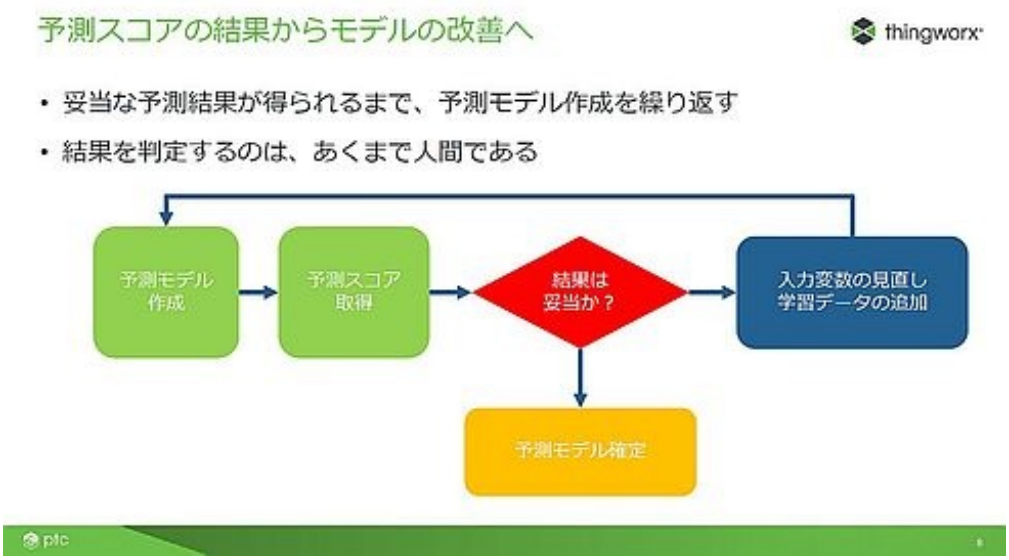

図7　予測スコアの結果からモデルの改善へ（クリックで拡大）

前編でも触れたが、機械学習による予測システムは、各企業で重要なパートを担うことが多くなかなか実例を紹介することが難しい。そこでPTCは、パートナー企業と共同でThingWorx Analyticsのデモ装置を開発した。

図8は、フローサーブ（Flowserve）のポンプ機器の故障を予測するデモ装置である。ポンプはフローサーブ、センサーはNI（National Instruments）、アプリケーションを動作させるエッジサーバ機器はHPE（HP Enterprise）、そしてソフトウェアはPTCが提供することで実現した。

図8　フローサーブのポンプ機器の故障を予測するデモ装置（クリックで拡大）

もともと構築されていたIoTによるポンプ機器の遠隔監視サービスアプリケーションに、ThingWorx Analyticsによる予測機能を付加させている。これによりポンプ機器の故障原因、及び3つの重要な部品について故障までの日数を予測できる。なお、フローサーブより提供されたデータに基づいて、あらかじめThingWorx Analyticsで作成した予測モデルを利用している

。

デモ装置では、写真の右側に見える赤いバルブを手動で閉めて異常状態を発生させることで、ポンプの故障と似たような状況をセンサーに認識させて推定される故障原因および部品の1つの故障までの予測日数を表示している（図9）。

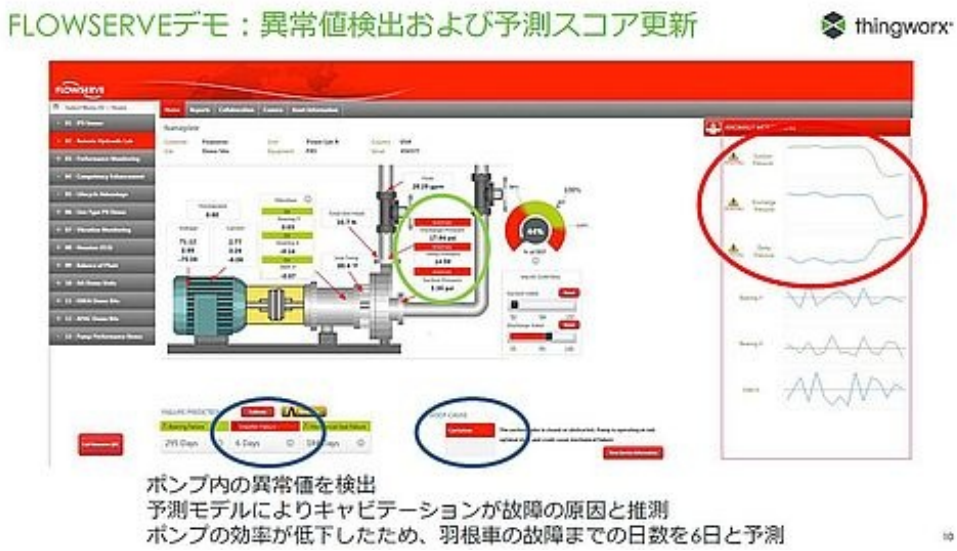

図9　デモ装置による異常値検出および予測スコアの更新（クリックで拡大）

このデモにもあるように、機械学習を利用した機器の故障予測機能を実現するためには、機器メーカーとツールメーカーによる連携だけではなく、センサーメーカーやサーバメーカーとの連携も必要となる。

6. あれこれ考えずに、まずはやってみよう

機械学習は、あれこれと手間が掛かるように感じられるかもしれないが、まずはやってみるのが肝心である。結果が判定しやすく効果が大きいテーマに絞りデータを集めたら、手を動かして実際にやってみることで、どのくらいの手間がかかり、どんな成果が得られるのか知見を得ることこそが財産となる。

PTCでは、単にソフトウェアツールを提供するだけでは無く、機械学習を使ったデータの分析／予測について、データの編集や実証実験など立ち上げに必要となる技術支援を有償サポートで用意しているので、社内にデータ分析のエンジニアが不足している場合は検討していただければと思う。

前編、後編の2回にわけて機械学習の概要を紹介した。AIという実体の分かりにくい言葉から、より具体的なイメージを描いていただければ幸いである。そして、経験と勘だけではないデータの力を得るためにも、機械学習によるデータ分析／予測に取り組む企業が増えることを切に願う。

筆者プロフィール

西 啓（にし あきら） PTCジャパン 製品技術事業部 IoT/Manufacturing技術本部 シニアIoTプリセールススペシャリスト

2015年9月にPTCジャパンに入社し、現職。日本におけるIoTに導入される機械学習、ARといった新技術の紹介、提案を実践している。さまざまなパートナー企業との関係を構築し、エコシステムによるビジネス拡大を画策中

・PTCジャパン

https://www.ptc.com/ja/

関連記事

AIと機械学習とディープラーニングは何が違うのか
技術開発の進展により加速度的に進化しているAI（人工知能）。このAIという言葉とともに語られているのが、機械学習やディープラーニングだ。AIと機械学習、そしてディープラーニングの違いとは何なのか。

世界を変えるAI技術「ディープラーニング」が製造業にもたらすインパクト
人工知能やディープラーニングといった言葉が注目を集めていますが、それはITの世界だけにとどまるものではなく、製造業においても導入・検討されています。製造業にとって人工知能やディープラーニングがどのようなインパクトをもたらすか、解説します。

ディープラーニングの事業活用を可能にする「ジェネラリスト」の重要性
AI技術として注目を集めるディープラーニング。ディープラーニングへの取り組みを進めていく上で必要とされる人材には「エンジニア」の他に「ジェネラリスト」も必要だ。本稿では、ディープラーニングの「ジェネラリスト」に何が求められるかについて解説する。

芽吹くか「組み込みAI」
第3次ブームを迎えたAI（人工知能）。製造業にとっても重要な要素技術になっていくことは確実だ。2017年からは、このAIを製品にいかにして組み込むかが大きな課題になりそうだ。

PTCが進めるフィジカルとデジタルの融合、その時「IoTは次世代のPLMになる」
PTCの年次ユーザーカンファレンス「LIVEWORX 2017」の基調講演に同社社長兼CEOのジェームズ・E・ヘプルマン氏が登壇。「PTCの役割は、革新を生み出すフィジカルとデジタルの融合の推進にある」と語るとともに、「IoTは次世代のPLMになる」と訴えた。

「フィジカル」と「デジタル」が融合するIoT時代、PTCはオープン化を加速する
PTCジャパンのユーザーイベント「PTC Forum Japan 2016」の基調講演に米国本社PTCの社長兼CEOを務めるジェームズ・E・ヘプルマン氏が登壇。「IoT時代のモノの新しい見方～現実世界とデジタル世界の収束」と題して、IoTプラットフォーム「ThingWorx」を中核とする同社の事業戦略を説明した。