品質・稼働ビッグデータを用いた異常検知についての提案

Suggestion of anomaly detection with industrial big data

原田 奈弥 [1*]　本村陽一 [1,2]
Nami HARADA[1]　Yoichi MOTOMURA[1,2]

[1] 産業技術総合研究所
[1] ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY
[2] 東京工業大学
[2] Tokyo Institute of Technology

Abstract: These days in almost all manufactual companies in Japan, A lot of data are collected frequently everytime. These data has much useful information about the quality and operation of their production process lines. However, it is difficult for us to use the data well. these data has so many dimensions that it is hard to understand instantly what is wrong with their production processes and decide what to do. In this paper, I suggest a way to understand the condition of their production process lines easily with some knowledge of their masters to help us decide what to do to improve their quality or operation.

1　はじめに

産業の国際化と情報技術の発展により、世界中がサプライチェーンとしてつながり、膨大なデータがリアルタイムで収集･蓄積することが可能になった。世界中から日々、刻々と生み出される品質や稼働に関する莫大なデータの、不具合発生防止のような品質管理や工程管理への活用は、産業の競争力やブランド力の維持向上の好機である。様々なデータを組み合わせ、データ間の関係や変動からビジネスにおける問題点の発見や課題解決などを行うことは、ビッグデータ活用と呼ばれ、工業をはじめとする様々な産業で切望されている。工業では、例えば不良品の流失防止のため、膨大なデータから不具合発生傾向を速やかに捉え、グローバルにまたがるサプライチェーンに対して働きかけを行いたい。人工知能の活用により、莫大なデータを活用していかにして不具合傾向を効率的に見付けるのかについて、本研究では取り組んだ。その取組み事例とその実践方法を以降に詳しく述べる。

2　問題提起

工場内の稼働や品質におけるデータ活用のために、

- 膨大なデータの情報圧縮や可視化による現状把握
- 品質や稼働に関する因果関係の推論を得ることによる固有技術の向上や知識の獲得
- 現在の固有技術やノウハウの織り込み
- 高次元データによる次元の呪いや過学習への対策

を行う必要がある。異なる工程間の品質や稼働に関するデータの活用は、予防保全などの品質や稼働の向上のために必要不可欠だが、データの変数が多く、原始的なクロス集計や管理図による管理は非効率的で、非現実的である。不具合の流出防止のためには問題の早期発見・早期対策が不可欠だからだ。ビッグデータ活用において、膨大な変数で構成されるデータから必要な情報の抽出と可視化が、人工知能の活用による業務の効率化の第一歩である。

また、日本の製造業の成長の歴史は、実験データや清算工程内データを用いた SQC 活動を通して地道に固有技術の知見やノウハウを獲得してきた結果である。これからも競争力を維持向上し、成長を続けるには、今まで蓄積されたノウハウや知見を活用しながら、その製品や生産工程に関する固有技術や知識の更なる獲得ができるビッグデータの活用を目指すことが必要である。

日々の生産の中で確率的に発生する品質や稼働についての問題を扱う上で、データ科学からのアプローチは有用性が高いが、データ科学の観点からビッグデータ活用を考えた場合に考慮すべき点は、データが高次元であることによる過学習である。高次元データで数

*連絡先：産業技術総合研究所
〒 135-0064 東京都江東区青海 2-4-7
E-mail: nami.harada@aist.go.jp

理モデリングによる統計的因果推論を行い、パラメータの適切な推定値を得て、工程内の品質や稼働に対して働きかけを行う上で、次元の呪いや過学習を考慮する必要がある。

工程内状態監視や予防保全に対する取り組みの先行事例として、IBM のアナコンダ、NEC のインバリアント分析などで、高度な機械学習を用いた工程内の異常検知の手法が提案されている。これらの従来のアプローチは、データ空間の中で正常値と異常値の識別を学習する、識別モデルの構築による異常検知を行ってきた。

本研究では、工程内不具合未然防止のためのビッグデータ活用の第一歩として、簡易な機械学習手法を用いて、該当の工程や作業に関する固有の技術や知識を、人工知能に対して対話的に織り込みながら、高次元データから品質や工程に関する情報を圧縮して抽出し、可視化することを目標とする。そのため、本研究では、工程内不具合に対して生成モデルからアプローチする。生成モデルを構築することで、不具合発生のメカニズムについて統計的な推測を出力として得て、工程や品質の管理のために取るべき行動を考えたり変えたりすることを助けたい。さらに、工程固有の暗黙知として知られている事象を対話的に人工知能に織り込み、人工知能の中での機械学習の結果に、これまでに生産現場で蓄積されたノウハウや知見を組み合わせることで、現実的な行動変容を効率よく促すためのプロセスを考えた。このような人工知能の実装・活用によって、データが観測される現場の人も納得できる、これまで蓄積された暗黙知と矛盾しない人工知能と共に品質や稼働を管理し、改善を促すことを目指した。

3 提案

工程内のビッグデータを用いて、簡易な機械学習手法を用いて品質や稼働に関する統計的因果推論を行う上で、決定木の手法は有用である。決定木とは、一つのカテゴリカルな目的変数に対して、説明変数が目的変数を変化させる要因を、ベイズの公式を用いて予測し、予測関数を求める手法である。決定木とは、目的変数のモードについて、説明変数となるパラメータをエッジとして、パラメータの状態別に目的変数がとる可能性が高いモード毎に枝を出す、木構造のモデルを学習する機械学習手法である。そのような構造をベイズの公式に基づき、学習・推測する。木構造で出力することで、目的変数のモードと説明変数の関係について統計的な推測が得られる。目的変数のモードを説明変数から予測することも可能である。

工程内の品質や稼働管理に関するビッグデータに対して決定木を用いる有用性は、説明変数と目的変数の関係についての統計的な因果推論を、決定木の出力の木構造から得られることである。この決定木の構造から、該当の工程の品質や稼働に関する固有技術を得ることができる。

本研究では、この決定木の出力結果から、工程内で観測された各パラメータの不具合発生確率への影響を確率によって評価し、不具合発生確率を工程内でのリスク管理指標として求め、この工程内リスク管理指標を用いて工程内の品質や稼働を管理することを提案する。具体的には、図 1 のような決定木の出力を得たときに、不具合モードの発生のしやすさの確率に、各説明変数となるパラメータに対して、そのパラメータの扱いにくさやその工程での不具合の品質についての重要性などの固有の知見を基に、重みを付けてリスク管理指標とすることを提案する。

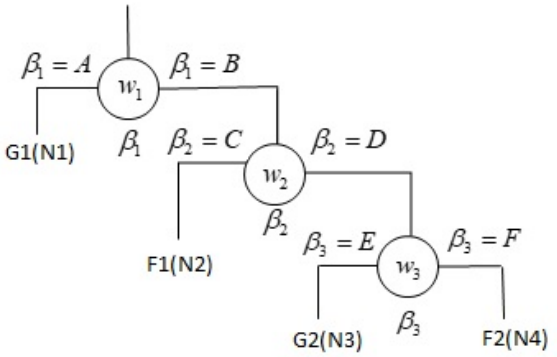

図 1: 決定木の事例

この決定木の表記の、$\beta_1 \sim \beta_3$ が工程内の生産条件などのパラメータで、それによって、品質の良品・不良品のモードが変わるとする。良品が、製品種類などのモードによって G1、G2 で、不良品が現象などのモードによって F1、F2 とする。G1、F1、G2、F2 の観測数がそれぞれ、N_1, N_2, N_3, N_4

$$N_1 + N_2 + N_3 + N_4 = N_T \tag{1}$$

とする。

このとき、不具合発生傾向 T(F) は、

$$T(F) = P(\beta_1 = B, \beta_2 = C) + P(\beta_1 = B, \beta_2 = D, \beta_3 = F) \tag{2}$$

$$\begin{aligned} P(\beta_1 = B, \beta_2 = C) &= P(\beta_2 = C | \beta_1 = B) \\ &= \frac{P(\beta_2 = C \cap \beta_1 = B)}{P(\beta_1 = B)} \\ &= \frac{\frac{N_2}{N_2+N_3+N_4}}{\frac{N_2+N_3+N_4}{N_T}} \\ &= \frac{N_2 \times N_T}{(N_T - N_1)^2} \end{aligned} \tag{3}$$

$$
\begin{aligned}
&P(\beta_1 = B, \beta_2 = D, \beta_3 = F) \\
&= P(\beta_3 | \beta_1 = B, \beta_2 = D) \\
&= \frac{P(\beta_3 = F \cap \{\beta_1 = B, \beta_2 = D\})}{P(\beta_1 = B, \beta_2 = D)} \\
&= \frac{\frac{N_4}{N_3+N_4}}{P(\beta_2 = D | \beta_1 = B)} \\
&= \frac{\frac{N_4}{N_3+N_4}}{\frac{N_3+N_4}{N_2+N_3+N_4}} \\
&= \frac{N_4(N_T - N_1)}{(N_3+N_4)^2}
\end{aligned} \tag{4}
$$

よって、

$$
T(F) = \frac{N_2 \times N_T}{(N_T - N_1)^2} + \frac{N_4(N_T - N_1)}{(N_3+N_4)^2} \tag{5}
$$

で、この工程内での不具合発生傾向$T(F)$が計算できた。

さらに、工程内パラメータ $\beta_1 \sim \beta_3$ に関して、工程への介入の難しさやコスト、品質不具合などの重要性に応じて、この式 (2) の工程内リスク管理指標に重み w_1〜w_3 をかけることで、この工程内リスク管理指標に対して、工程に関する特有のノウハウや暗黙知を織り込むことができる。このときの工程内リスク管理指標 RMI(Risk Management Indicator) を、

$$
\begin{aligned}
RMI \quad = \quad & w_2 P(\beta_1 = B, \beta_2 = C) \\
& + w_3 P(\beta_1 = B, \beta_2 = D, \beta_3 = F)
\end{aligned} \tag{6}
$$

と定義した。ただし、

$$
0 < w_1, w_2, w_3 \leq 1 \tag{7}
$$

で、

$$
w_1 = w_2 = w_3 = 1 \tag{8}
$$

のとき、式 (6) は

$$
T(F) = RMI \tag{9}
$$

となる。つまり、重み w_1〜w_3 は、その工程内のパラメータをどれだけ軽視できるかを、恣意的に設定するパラメータとみなすことができる。部品の交換などの工程への介入が易しいことが予め分かっているときは、重み w_1〜w_3 の値を小さくすることで、現在の工程固有のノウハウを織り込んだ工程管理が可能となる。式 (2) や式 (6) の指標を、その工程内での不具合の起こりやすさを示す値として管理することを本研究では提案する。試作などの際に、重み w_1〜w_3 の値を変えながら品質などの試験を行うことで、実際の工程内の品質や稼働とこの提案している工程内リスク管理指標の妥当性や信頼性を評価することも可能である。

4 事例

本研究では、製造工程の品質データとは異なるが、タイタニック号の生存者と、その性別と年齢、宿泊していた船室の等級のデータを用いた。性別と年齢、船室の等級から、$T(F)$ は死者の比率と考えてモデリングを行った。なお、本研究では決定木は統計解析ソフト R Ver. 3.3.1 の、{rpart} パッケージを用いた。 このデータには、総勢 2201 人の乗客と船員について、それぞれの名前と、船室の等級が 1〜3 等と船員の 4 クラス、性別が男性・女性の 2 クラス、年齢が大人と子供の 2 クラスと、生存について生存か死亡かの 2 クラスのデータが記録されている。

モデリングは、式 (2) の式の、乗客・乗員の全体の死亡者率を $P(F)$ として、死亡する要因を、船室の等級と性別、年齢で決定木でモデリングを行った。その結果が、図 2 である。この図 2 の、分岐の一番最後に

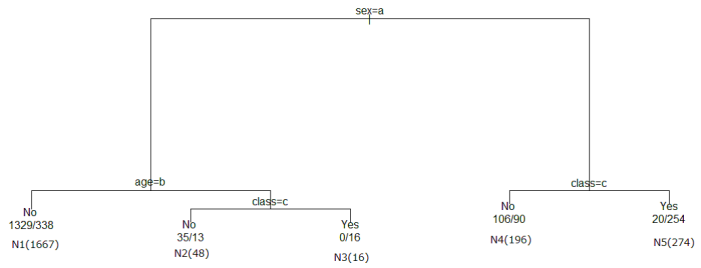

図 2: タイタニックの事例

割り当てられた観測の人数を左から順に、$N1$〜$N5$ とした。ここで、

$$
N1 + N2 + \cdots + N5 = N_{TT} = 2201 \tag{10}
$$

とする。図 2 の、1 番上の分岐の性別が a(男性) が左で、女性が右となっている。船室の等級 (class) が c(3 等級) であるかどうかと、年齢の、生存との間の統計的因果関係が推測できる。生存は、一番下の葉が、Yes が生存、No が死亡を示している。

この中で、死亡 (No) となっている現象が観測された人は、図 2 の、$N1$ の男性 (sex=a) で大人 (age=b) の人が 1329 人と、$N2$ の男性 (sex=a) で子ども (age≠b) で船室が 3 等級 (class=c) の人が 35 人、$N4$ の女性 (sex≠a) で船室が 3 等級 (class=c) の人が 106 人となっている。このデータの、式 (2) の $T(F)$ は死者発生傾向と

なり、

$$
\begin{aligned}
&T(F) = T(\text{life} = \text{No}) \\
&= P(\text{sex=a, age=b}) \\
&\quad + P(\text{sex=a, age} \neq \text{b, class=c}) \\
&\quad + P(\text{sex} \neq \text{a, class=c}) \qquad (11) \\
&P(\text{sex=a, age=b}) = P(\text{age=b}|\text{sex=a}) \\
&\quad = \frac{P(\text{sex=a} \cap \text{age=b})}{P(\text{sex=a})} \\
&\quad = \frac{N1}{N1+N2+N3} \qquad (12) \\
&P(\text{sex=a, age} \neq \text{b, class=c}) \\
&\quad = P(\text{class=c}|\text{sex=a, age} \neq b) \\
&\quad = \frac{P(\text{class=c} \cap \{\text{sex=a, age} \neq \text{b}\})}{P(\text{sex=a, age} \neq \text{b})} \\
&\quad = \frac{P(\text{class=c} \cap \{\text{sex=a, age} \neq \text{b}\})}{\frac{P(\text{age}\neq\text{b}\cap\text{sex=a})}{P(\text{sex=a})}} \\
&\quad = \frac{N2}{N1+N2+N3} \qquad (13)
\end{aligned}
$$

$\because$

$$
\begin{aligned}
&P(\text{class=c} \cap \{\text{sex=a, age} \neq \text{b}\}) \\
&\quad = P(\text{sex=a, age} \neq \text{b}) \\
&\quad\ \times P(\text{class=c}|\text{sex=a, age} \neq \text{b}) \\
&\quad = P(\text{age} \neq \text{b}|\text{sex=a}) \\
&\quad\ \times \frac{P(\text{class=c} \cap \{\text{sex=a, age} \neq \text{b}\})}{P(\text{sex=a, age} \neq \text{b})} \\
&\quad = \frac{N2+N3}{N1+N2+N3} \\
&\quad\ \times \frac{P(N2)}{P(\text{age} \neq \text{b}|\text{sex=a})} \\
&\quad = \frac{N2+N3}{N1+N2+N3} \times \frac{N2}{N2+N3} \\
&\quad = \frac{N2}{N1+N2+N3} \\
&P(\text{age} \neq b|\text{sex=a} = \frac{P(\text{age} \neq \text{b} \cap \text{sex=a})}{P(\text{sex=a})} \\
&\quad = \frac{N2+N3}{N1+N2+N3} \\
&P(\text{sex} \neq a \cap \text{class=c}) = \frac{P(\text{sex} \neq a, \text{class=c})}{P(\text{sex} \neq a)} \\
&\quad = \frac{N4}{N4+N5} \qquad (14)
\end{aligned}
$$

式 (11)、式 (12)、式 (13)、式 (14) より、このデータの $T(F)$ は、

$$
T(F) \quad = \quad \frac{N1+N2}{N1+N2+N3} + \frac{N4}{N4+N5} \qquad (15)
$$

となる。式 (15) と図 2 より、この事件の $T(F)$ は、1.41 となった。

5 考察・まとめ

本研究では、例えば量産工程内の不具合のような望ましくない事象について、望ましくない事象の発生メカニズムについて、決定木による生成モデルの構築によって推測を得ながら、その望ましくない事象の工程内での起こりやすさを、発生する条件付確率に基づきモニタリングするプロセスを、タイタニックの事例とデータを用いて実証した。本研究では、製品の品質状況 (良品/不良品) という一変数の目的変数について、統計的な因果推論を階層的に行うために、決定木を用いた。

通常決定木を生成モデルとして用いる場合、データ間の統計的な因果推論のグラフィカルな記述に用いられる。つまり、既に観測されたデータの変数間に、統計的に考えられる関係を学習して記述することに用いられてきたが、本研究にて提案する手法では、それを、工程内のノウハウや固有の知見と合わせながら、これから先の不具合発生の予測に用いる手法を提案しているところに新規性がある。

本研究で提案した不具合発生傾向 $T(F)$ は、決定木の出力に基づく確率から計算しているが、1 を超える可能性がある。なぜなら、条件付発生確率から $T(F)$ を求めているので、決定木の上位の枝の起こりやすさが反映されるからである。具体的には、図 2 において、死ぬと推測されるノードは $N1, N2, N4$ なので、死ぬと予測されるノードに属する人の割合は $(N1+N2+N4)/N_{TT}$ だが、上位の枝の、性別や年齢、船室の等級に割り当てられる事前確率を考慮した、$N1, N2, N4$ への割り当てられる確率をそれぞれ求め、その和を $T(F)$ としているため、1 を超えることがある。

本研究の提案手法の長所は、

- 工程内の現在の固有の知識を活かした人工知能の活用
- 高次元データの、工程内危険指標の 1 次元への圧縮と可視化
- 決定木による、工程内データの統計的因果推論を行うことによる固有技術や知識の継続的な獲得
- シンプルな仮定やモデリングにより、様々な生産工程に柔軟に対応
- 変数の重みパラメータによる対話的な人工知能の活用

である。

日本のものづくりの品質は、工業のみならず様々な分野で世界的な信頼を得るほど技術やノウハウがすでに高いレベルにある。日本のものづくりにおける人工知能の適用の課題の一つは、いかにして現在の品質や稼働を守り、ノウハウを活かすかではないだろうか。本

研究では、決定木の中の各説明変数に、人為的に決定できる重みパラメータを設定することで、現在のノウハウを織り込むことができた。

工程内の品質や稼働に関する高次元データを、不具合発生の確率に基づく工程内リスクとして 1 次元の指標にすることで、品質や稼働に関する傾向として理解・管理をしやすくすることができた。決定木のモデリングに基づく条件付確率という、シンプルな計算プロセスであるため、様々な生産プロセスの中で柔軟に対応できるのではないかと期待している。不具合発生確率 $P(F)$ を式 (2) で求められるので、不具合のモードが複数ある工程や、様々な仕向やオプションの製品が流れるような、多品種生産工程にも適用できる。あるいは、不具合のモードの中でも特にコストやリスクが高い不具合モードを選んで検証することなど、自由に柔軟に適用できることも本提案の長所である。

さらに、決定木による工程内の観測データと品質や稼働に関する統計的な因果推論を得ることで、機械学習や人工知能だけでなく、工程内の品質や稼働についても固有の知識や技術、ノウハウを今後も継続して得ることができる。人工知能の、日本のものづくりの信頼やブランド力、競争力の維持向上への貢献は、私の切なる願いである。

重みパラメータを対話的に操作することで、リスク管理指標をより適切なものにすることができる。生産設備の変更など、工程の変化点がある度に、重みパラメータの適切な値を、工程内の特有の条件を織り込み、分かりやすく柔軟に変更することができることも本手法の長所である。人と相互理解をしながら、相互に作用しあうことで人とともに成長できる人工知能こそ、生産工程における活用が日本の今後のものづくりの発展に大きく貢献できると考えているからだ。

6　今後の課題

今後の課題は傾向管理における本手法の有用性の検討である。ビッグデータ活用による工程管理の実務上最も重要なニーズの一つは、傾向的に徐々に発生しやすくなっていく不具合などの問題について、大きな問題の発生を未然に予測して防止することである。本研究で事例として用いたタイタニックのデータは、事故が起きたある 1 点の観測時点で観測されたデータで、経時的に反復・継続して観測したデータではない。本研究にて提案した手法が、時間の変化とともに徐々に変化していくものであるのか、継続的に観測していくことで品質や稼働に関する問題を未然に防止することに有用であるかどうかを今後検証する必要がある。

工業の量産工程の品質管理への応用に向けて、上記のように同じ工程内で傾向管理するために繰り返し測定されたデータで、特に、生産時の加熱温度や締め付けトルクのような、工程内で人が操作可能な生産条件を品質の変動要因として割り当てて観測する必要がある。提案手法の有用性の検証のためには、式 (2) の不具合発生確率 $T(F)$ や工程内リスク管理指標の式 (6) の RMI が生産条件を変えることでどのように変化し、$T(F)$ や RMI の数値の改善と共に品質や稼働も本当によくなっているのかを検証する必要があるが、本研究で使ったタイタニックの事例では、死亡を説明する変数として当てはめたデータは第 3 者が操作することができない。

本手法を実際の活用に際しては、重みパラメータの設定の高い恣意性のため、対話的な活用が可能となった一方で、その恣意性がまた本手法活用のハードルとなることが考えられる。この課題をクリアするためには、各重みパラメータが関わる工程についてのリスクやコストを評価し、重みパラメータに反映していく手法の標準化が今後望まれる。

本研究では、統計的因果推論に決定木を用いた。これは、フリーシェアで広く使われている R で使える、シンプルで理解しやすい手法として選んだが、モデル選択やデータの形式など、手法に関する理論的なアプローチも継続して行っていきたい。決定木のような、データ内の変数間の統計的な因果推論を学習しグラフィカルに出力する機械学習の手法は決定木の他にも、ベイジアンネットワークや構造方程式モデリングなどもある。また、製造業の稼働に関するデータは、通常の量産工程では正例に比べ、負例が圧倒的に少ない、アンバランスデータであることが通常である。そのようなデータで、階層的な変数間の構造を安定して推定するためには、決定木は適切かどうか、どのような課題があり、どうやって解決していくのか、今後考えていきたい。

今後も製造業のサプライチェーンの国際化やセンサリングなどの科学技術の発展が続けば、観測・データ収集できる工程内の変数が増え、試作段階では、データの観測数よりも観測される変数の方が多い、高次元データとなることが考えられる。このようなデータの性質も踏まえ、学習手法の違いを比較し、適切な手法を選択することを今後考えたい。

また、本研究では、製品の品質という一変数を目的としたモデリングのため、決定木を用いた。産業への応用を考えると、最小コストで最良の品質といった、多目的最適化の問題も多くみられる。このような、応用現場のニーズに応えるモデルの高度化も今後の課題である。

また、本手法をサービスや、快適さや安心などの定性的、もしくは主観的な性質が強い目的変数に適用する場合は、上記の重みパラメータに加え、目的変数を定量的にどう定義するのかを検討する必要がある。日本品質として世界に知られるものづくりだけでなく、

ホスピタリティの高い、おもてなしの精神に基づく日本のサービス業も日本の競争力を支える重要な一分野であり、この競争力の継続的な維持向上には人工知能の活用が有効だと考えている。工程間のデータの統合によるビッグデータや人工知能の活用には、定性的や主観的な情報の定量化が必要であり、この定性的・主観的な指標の定量的な評価が、本提案手法の活用含め、サービス業における人工知能の活用の肝となるのではないか。

謝辞

本研究は NEDO 委託事業「人間と相互理解できる次世代人工知能技術の研究開発」の支援を受けて行いました。

参考文献

[1] P. Hendricks: Package ‘titanic’, *CRAN*, (2015)

[2] T. Therneau, B. Atkinthon and B. Ripley: Package ‘rpart’, *CRAN*, (2017)

[3] 荒木雅弘: フリーソフトではじめる機械学習入門, 森北出版 (2014)

[4] 井手剛, 杉山将: 機械学習プロフェッショナルシリーズ 異常検知と変化検知, 講談社 (2015)

[5] 本村陽一: ベイジアンネットワークによる確率的人間行動モデリング, 電気通信大学大学院 電気通信学研究科 学位論文 (2008)