特集記事 • 6
鉄鋼業におけるAI・IoT技術の最前線

材料工学へのデータサイエンス手法の適用

Applicability of Data Science Method on Materials Science and Engineering

名古屋大学 大学院工学研究科 教授 小山敏幸 Toshiyuki Koyama
名古屋大学 大学院工学研究科 准教授 塚田祐貴 Yuhki Tsukada

1 はじめに

近年、材料工学とデータサイエンスとの融合が国内外で活発化している。いわゆるマテリアルズ・インフォマティクスやマテリアルズ・インテグレーション（以降、MIと記す）と呼ばれる分野である[1-5]。その中でも日本鉄鋼協会は、材料の組織と特性部会「鉄鋼インフォマティクス研究会（名古屋大学 足立吉隆 主査）」を皮切りに、比較的早期から当該分野の議論をスタートさせている点を、まず強調しておきたい[6]。

さて最近、人口知能（以降、AIと記す）や機械学習が、材料の分野で注目されているわけだが、本稿では、一歩下がって、材料工学へのデータサインス手法の適用について考えてみたい。著者はAIや機械学習は、材料工学の発展に有用と考えているが、もちろん万能とは考えていない。スタンスとしては、"多くの新ツールや、データそのものに対する扱いで有益な考え方が出てきたので、使わないのは損である"といった立場である。以下、一般的な内容から、注目すべき解析法、当該分野に関する俯瞰的内容等について、やや私見も含めて順次説明させていただきたい。

2 インフォマティクスの定義

インフォマティクスの定義から考えてみよう。インフォマティクスをウィキペディアで検索すると、種々の定義がでてくる（ウィキペディアは玉石混合で必ずしも正しいわけではないが、情報量が豊富である利点はある）。結局、バイオインフォマティクスやケミインフォマティクスの流れで、用語「マテリアルズインフォマティクス」が選択されたと類推できる[7,8]。ここでは議論の対象をクリアにするために、材料開発研究の観点からインフォマティクスを定義すると、現在、材料工学で議論されているインフォマティクスとは、「探索や最適化の圧倒的高効率化を実現する方法論（知りたいことに、迅速・効率的にたどり着く方法論）」としてよいであろう。さて、インフォマティクスが役立つためには、当然ながら、探索される対象のデータが必要であり、対象のデータが巨大であればあるほど、インフォマティクスの恩恵は大きくなる。いわゆるビッグデータ[9]が存在してはじめて、インフォマティクスで培われたツールは威力を発揮する。以上から、インフォマティクスが活躍できる必要条件は以下の4つにまとめられる。

(1) 実験データの大容量化、高精度化、および共有化
(2) 理論とシミュレーションの高度化（各種のデータ間を縦横に結ぶパーツの充実）
(3) 機械学習の高度化・実用化（数学的手法の整備、ツールの汎用化、計算機能力の向上）
(4) 人材育成

材料工学の進展における近年の特徴は、各種の計測・分析機器の革新、および理論・シミュレーション手法の高度化であろう。つまりこれら解析手法の進展に伴い、上記4条件が材料学分野で整い始めたことが、現在のMI隆盛の背景にある。注意すべき点は、材料分野は、気象やインターネット分野ほどの日常的なビッグデータ取得に未だ至っていない点である。つまり材料分野全般は、先行して進んでいる分野の手法が、そのまますぐに役立つ環境にはない。しかし、先人のたゆまぬ努力によって大量のデータが入手できる（もしくは生成できる）分野は存在するので、そのような分野から、各種の展開が始まっているのが実状である。したがって、ビッグデータには至らないが、そこそこのデータ数がある場合のデータ活用法や、実験・理論・シミュレーションを駆使した補間によるデータ数の短期的拡大、また比較的ビッグデータを得やすい間接データ（画像や音など）の活用などが、直近の実用的課題であろう。

3 データの区別

機械学習の流行によって、"ベイズ"という用語に頻繁に出くわすようになった。事前確率と、実験・理論から得られるデータに基づき、事後確率を最大化するベイズ推定[10]は、基本的に材料学に相性のよい考え方である（ベイズ推定は、結局、[仮設⇒実験⇒検証⇒仮説の修正⇒…]の繰り返しになるので、これは試行錯誤の色彩の強い材料学の研究アプローチそのものに近い）。一方、データの統計解析手法には、もともと多変量解析や実験計画法などがあり[11,12]、これらも工学における有益な手法である。これら従来の統計解析と、ベイズ推定に基づく統計解析の相違を理解するには、対象とするデータを、以下の2種類に分類するとわかりやすい（なお以下の用語は、著者の造語であるので注意されたい）。

(1) 有限孤立型データ　⇒　解析手法は、多変量解析や実験計画法などの従来の統計解析
(2) 無限開放型データ　⇒　解析手法は、ベイズ推定等を基礎とした機械学習

(1) はデータ数が有限で、新しいデータが追加されない前提にて統計量を求める場合である。当然ながら、誰がやっても同じ結果（例えば、平均や分散など）が得られるので、普遍性が高く結果の信頼度も大きい。一方、(2) は時々刻々とデータが降ってくるような環境にて（この意味で、データ開放型と記した）、統計量を算出するような場合に相当する。特にビッグデータ環境を想定すると、必然的に、ベイズ推定の方法論を取らざるを得ない。例えば平均値の計算で、毎回、全ビッグデータを足し合わせて平均を計算する手順は明らかに非現実的である。新しいデータが付け加わるたびに、その都度、事前の平均値を修正する手順の方が現実的であろう。ただし事前の平均値と、直近に使用した測定データに依存して、事後の平均値が変わり得るので普遍性は損なわれる。このため統計学の分野で、ベイズ推定を普遍的な学問として認知するかどうかの論争が長く続けられたが、無限開放型のビッグデータを扱う時代に入り、実用的な観点から、その圧倒的有用性が認知された[13]。以上から (1) と (2) の相違を理解して、解析手法を使い分けることが重要であろう。

4 データ同化

機械学習手法の種類は膨大であり、かつ一つの手法にも数多くの派生版が存在する[14-16]。ここでは、その中でも材料工学において特に重要な概念・手法となり得る「データ同化[17,18]」について、やや詳しく取り上げる。データ同化は、Data Assimilationの訳であり、Assimilationは、異なる文化が融合する意味に用いられる単語である。材料研究を対象に説明すると（やや狭義な説明であるが）、データ同化は、"実験データをシミュレーションに融合させ、シミュレーションの予測精度を向上させる方法論"である。シミュレーションと理論は演繹的アプローチに属し、一方、実験・測定は帰納的アプローチに属す。両者は相補的な関係となる場合が多く、データ同化は、まさに材料研究における理想的アプローチの最終形態といえる。

さて、データ同化には逐次型と非逐次型があり、逐次型データ同化は、新しい実験データが加わるたびにシミュレーションに含まれる各種のパラメータや初期・境界条件等をより適切な値に修正する方法である。一方、非逐次型は、得られた実験データを固定し、その条件下で、シミュレーションに含まれる各種のパラメータ等を最適な値に修正する方法である。前者は、無限開放型データを対象としたデータ同化となり、カルマンフィルタなどの各種のフィルタ理論がその代表的手法である。後者は有限孤立型データを対象としたデータ同化で、各種のフィルタ理論だけでなく、その代表的な手法に「アジョイント法」がある（なお有限孤立型データを対象とした解析は、通常の回帰・最適化問題に帰着できるので、通常の統計解析や機械学習を活用した各種の手法も活用できる）。

フィルタ理論もアジョイント法も、間違いなく材料学において極めて有益な手法であるが、この分野の通常の教科書において抽象的な説明が終始展開されるため、他の分野の研究者・技術者には、非常に学習しにくいというのが著者の偽らざる感想である。フィルタ理論に関して、最もイメージが沸きやすい学習手順は、「粒子フィルタ[17]」を対象に、まず「マルコフ連鎖モンテカルロ法[10,14,19]」を先に理解し、これを時系列で連続操業すると考えるとよいと思う。その後に、カルマンフィルタをはじめとする各種のフィルタ理論に進んだ方が理解しやすい（フィルタ理論の教科書は種類も多く、比較的容易に入手できるが、学ぶ順番により注意を払うべきであろう）。以下では、もう1つのデータ同化の手法であるアジョイント法[18]を詳しく取り上げたい。アジョイント法は、データ同化（特に気象予報）の分野で日常的に活用されている実用的な手法であるが、その理論に関する解説書は、フィルタ理論などの他のデータ同化の手法に比較してかなり少ない。当然ながら、材料技術者向けの解説は皆無といってよい。著者は、この理論は、鉄鋼材料学をはじめ、材料工学の全分野において、大きな可能性をもたらすと考えるので、以下において、材料研究者・技術者を念頭に、教科書的にアジョイント法の理論を説明させていただく。

4.1 アジョイント法の基礎

アジョイント法（4次元変分法とも呼ばれる）は、データ同化におけるパラメータ推定もしくは初期値推定に関する逆問題に主に活用され、順問題に微分方程式が使われるモデルに

おいて威力を発揮する方法論である[18]。非常に優れた手法であるが、随伴方程式を用いた数学的枠組みを巧みに利用する手法であるため、理論を直感的に理解することが難しい。以下では、通常の拡散方程式の数学体系で、濃度プロファイルの実験情報から、拡散係数Dの値（ここでは定数とする）、および時間ゼロにおける初期濃度プロファイル$c(x,\ 0)$、を推定する問題を取り上る。以下、具体的な定式化を説明する。

まず一次元（x方向）の拡散方程式は、

$$\frac{\partial c}{\partial t}-D\frac{\partial^2 c}{\partial x^2}=0$$

にて与えられる。濃度プロファイル$c(x,\ t)$は、位置xおよび時間tの関数である。また拡散係数Dを定数としたので、Dの時間依存$D(t)$を表す関係式として、

$$\frac{\partial D}{\partial t}=0$$

も考慮する（Dを時間tの関数とみなしている点に注意）。次にコスト関数Iを、

$$I\equiv\int_t\int_x J dxdt=\frac{1}{2}\int_t\int_x w(t)\{c(x,t)-c_{\text{obs.}}(x,t)\}^2 dxdt$$

と定義する。ここでJを、

$$J\equiv\frac{1}{2}w(t)\{c(x,t)-c_{\text{obs.}}(x,t)\}^2$$

とおいた（Jは拡散流束ではないので注意されたい）。$w(t)$は濃度プロファイルの測定値が存在する時間のみ$w(t)=1$で、他の時間は$w(t)=0$となる関数である。$c_{\text{obs.}}(x,\ t)$は濃度プロファイルの実験データで、$c(x,\ t)$が計算データである。コスト関数は、実験データと計算結果とのくい違いを表現する関数であれば、任意に定義してよい。ここでは差の二乗を時間および空間積分した量にて定義している。したがって、この問題は、コスト関数Iを最小化する$D(0)$と$c(x,\ 0)$を求める問題となる。

さて、求める対象は$D(0)$と$c(x,\ 0)$であるが、数値計算でこれらを求めるために必要な量は、$\partial I/\partial D(0)$と$\partial I/\partial D(x,0)$である。これらがわかれば、例えば通常の勾配法を用いて、$D(0)$と$c(x,\ 0)$を収束計算から求めることができる。実はアジョイント法は、$\partial I/\partial D(0)$と$\partial I/\partial c(x,0)$を巧妙かつ効率的に計算する手法に他ならない。

以下、数学的な準備を行う。まずラグランジュアンを、

$$\begin{aligned}L&=I+\int_t\int_x\lambda_c\left[D\frac{\partial^2 c}{\partial x^2}-\frac{\partial c}{\partial t}\right]dxdt-\int_t\lambda_D\left[\frac{\partial D}{\partial t}\right]dt\\&=\int_t\int_x J dxdt+\int_t\int_x\lambda_c\left[D\frac{\partial^2 c}{\partial x^2}-\frac{\partial c}{\partial t}\right]dxdt-\int_t\lambda_D\left[\frac{\partial D}{\partial t}\right]dt\end{aligned}\quad\cdots\cdots(1)$$

にて定義し（λ_cとλ_Dはラグランジュの未定乗数である）、この変分δLを計算する（Dが位置の関数ではない点に注意されたい）。

$$\begin{aligned}\delta L=&\int_t\int_x\left[\left(\frac{\partial J}{\partial c}\right)\delta c+\left(\frac{\partial J}{\partial D}\right)\delta D\right]dxdt\\&+\int_t\int_x\lambda_c\left[D\frac{\partial^2(\delta c)}{\partial x^2}-\frac{\partial(\delta c)}{\partial t}\right]dxdt\\&+\int_t\int_x\lambda_c\left[(\delta D)\frac{\partial^2 c}{\partial x^2}\right]dxdt-\int_t\lambda_D\left[\frac{\partial(\delta D)}{\partial t}\right]dt\\&+\int_t\int_x\delta\lambda_c\left[D\frac{\partial^2 c}{\partial x^2}-\frac{\partial c}{\partial t}\right]dxdt-\int_t\delta\lambda_D\left[\frac{\partial D}{\partial t}\right]dt\end{aligned}$$

右辺第二項と、［第三項＋第四項］は、部分積分を用いて、それぞれ以下のように変形できる。

$$\begin{aligned}&\int_t\int_x\lambda_c\left[D\frac{\partial^2(\delta c)}{\partial x^2}-\frac{\partial(\delta c)}{\partial t}\right]dxdt=\int_t\int_x\left\{D\left(\frac{\partial^2\lambda_c}{\partial x^2}\right)+\frac{\partial\lambda_c}{\partial t}\right\}(\delta c)dxdt\\&+\int_x\left[\lambda_c(x,0)\delta c(x,0)-\lambda_c(x,t_{\max})\delta c(x,t_{\max})\right]dx,\end{aligned}$$

$$\begin{aligned}&\int_t\int_x\lambda_c\left[(\delta D)\frac{\partial^2 c}{\partial x^2}\right]dxdt-\int_t\lambda_D\left[\frac{\partial(\delta D)}{\partial t}\right]dt\\&=\int_t\left\{\int_x\lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx+\frac{\partial\lambda_D}{\partial t}\right\}(\delta D)dt+\left[\lambda_D(0)\delta D(0)-\lambda_D(t_{\max})\delta D(t_{\max})\right]\end{aligned}$$

t_{max}は時間tの最大値（対象とするデータ範囲の最後の時間）である。これらをもとの式に代入して、

$$\begin{aligned}\delta L=&\int_t\int_x\left\{\left(\frac{\partial J}{\partial c}\right)+D\left(\frac{\partial^2\lambda_c}{\partial x^2}\right)+\frac{\partial\lambda_c}{\partial t}\right\}(\delta c)dxdt\\&+\int_t\left\{\left(\frac{\partial J}{\partial D}\right)+\int_x\lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx+\frac{\partial\lambda_D}{\partial t}\right\}(\delta D)dt\\&+\int_x\left[\lambda_c(x,0)\delta c(x,0)-\lambda_c(x,t_{\max})\delta c(x,t_{\max})\right]dx\\&+\left[\lambda_D(0)\delta D(0)-\lambda_D(t_{\max})\delta D(t_{\max})\right]\\&+\int_t\int_x\delta\lambda_c\left[D\frac{\partial^2 c}{\partial x^2}-\frac{\partial c}{\partial t}\right]dxdt-\int_t\delta\lambda_D\left[\frac{\partial D}{\partial t}\right]dt\end{aligned}$$

を得る。

さてここで、$\delta(L\text{-}I)=0$を要請しよう（つまり$\delta L=\delta I$）。$(L\text{-}I)$はもともと未定乗数が関与する項のみで表現される量であるので（式（1）を参照）、制約条件から$\delta(L\text{-}I)=0$は自明であろう。以上から、

$$\begin{aligned}\delta L=\delta I=&\int_t\int_x\left\{\left(\frac{\partial J}{\partial c}\right)+D\left(\frac{\partial^2\lambda_c}{\partial x^2}\right)+\frac{\partial\lambda_c}{\partial t}\right\}(\delta c)dxdt\\&+\int_t\left\{\left(\frac{\partial J}{\partial D}\right)+\int_x\lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx+\frac{\partial\lambda_D}{\partial t}\right\}(\delta D)dt\\&+\int_x\left[\lambda_c(x,0)\delta c(x,0)-\lambda_c(x,t_{\max})\delta c(x,t_{\max})\right]dx\\&+\left[\lambda_D(0)\delta D(0)-\lambda_D(t_{\max})\delta D(t_{\max})\right]\\&+\int_t\int_x\delta\lambda_c\left[D\frac{\partial^2 c}{\partial x^2}-\frac{\partial c}{\partial t}\right]dxdt-\int_t\delta\lambda_D\left[\frac{\partial D}{\partial t}\right]dt\end{aligned}\quad\cdots\cdots(2)$$

となる。ところでIは、$D(0)$と$c(x, 0)$が決まれば、その値が決まる関数であるので（$D(0)$と$c(x, 0)$が決まれば、通常の拡散シミュレーションで$c(x, t)$が決まる）、Iを$D(0)$と$c(x, 0)$のみの関数とみなすと、式(2)のδIは、$\delta c(x, 0)$と$\delta D(0)$が現れる項のみにて、

$$\delta I = \int_x \lambda_c(x,0)\delta c(x,0)\,dx + \lambda_D(0)\delta D(0) \quad \cdots\cdots (3)$$

と表記できなくてはならない。したがって、式(2)と式(3)より、

$$\int_t\int_x \left\{\left(\frac{\partial J}{\partial c}\right) + D\left(\frac{\partial^2 \lambda_c}{\partial x^2}\right) + \frac{\partial \lambda_c}{\partial t}\right\}(\delta c)\,dx\,dt$$
$$+\int_t \left\{\left(\frac{\partial J}{\partial D}\right) + \int_x \lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx + \frac{\partial \lambda_D}{\partial t}\right\}(\delta D)\,dt$$
$$+\int_x \left[-\lambda_c(x,t_{\max})\delta c(x,t_{\max})\right]dx + \left[-\lambda_D(t_{\max})\delta D(t_{\max})\right]$$
$$+\int_t\int_x \delta\lambda_c\left[D\frac{\partial^2 c}{\partial x^2} - \frac{\partial c}{\partial t}\right]dx\,dt - \int_t \delta\lambda_D\left[\frac{\partial D}{\partial t}\right]dt = 0$$

が成立し、関係式：

$$\frac{\partial \lambda_c}{\partial t} = -D\left(\frac{\partial^2 \lambda_c}{\partial x^2}\right) - \frac{\partial J}{\partial c},$$
$$\frac{\partial \lambda_D}{\partial t} = -\int_x \lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx - \left(\frac{\partial J}{\partial D}\right),$$
$$\lambda_c(x,t_{\max}) = 0, \quad \lambda_D(t_{\max}) = 0,$$
$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}, \quad \frac{\partial D}{\partial t} = 0 \quad \cdots\cdots (4)$$

が導かれる。さらに式(3)を、あらためて、全微分の形を意識して眺めると、

$$\lambda_c(x,0) = \frac{\partial I}{\partial c(x,0)}, \quad \lambda_D(0) = \frac{\partial I}{\partial D(0)}$$

であることがわかる（汎関数微分の観点できちんと定義すべきであるが、直感的な理解を優先している点を記しておく）。もともと知りたい量は、$\partial I/\partial D(0)$と$\partial I/\partial c(x, 0)$であったので、以上から、$\partial I/\partial D(0)$と$\partial I/\partial c(x, 0)$を求める問題が、時間0におけるラグランジュの未定乗数：$\lambda_c(0)$と$\lambda_D(x, 0)$を求める問題に置き換えられたことがわかる。

さて、あらためて式(4)を眺めると、λ_Dとλ_cに関する偏微分方程式（アジョイント方程式と呼ばれる）が、

$$\frac{\partial \lambda_c}{\partial t} = -D\left(\frac{\partial^2 \lambda_c}{\partial x^2}\right) - \frac{\partial J}{\partial c} \quad \Rightarrow \frac{\partial \lambda_c}{\partial(-t)} = D\left(\frac{\partial^2 \lambda_c}{\partial x^2}\right) + \frac{\partial J}{\partial c},$$
$$\frac{\partial \lambda_D}{\partial t} = -\int_x \lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx - \left(\frac{\partial J}{\partial D}\right) \Rightarrow \frac{\partial \lambda_D}{\partial(-t)} = \int_x \lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx + \left(\frac{\partial J}{\partial D}\right)$$

と書けることから、これらの微分方程式は時間を過去へさかのぼる微分方程式となっており、さらにこれらのアジョイント方程式の初期条件が、なんと、式(4)において、$\lambda_c(x, t_{\max}) = 0$および$\lambda_D(t_{\max}) = 0$にて、すでに与えられていることに驚く。つまり上記のアジョイント方程式を、$\lambda_c(x, t_{\max}) = 0$および$\lambda_D(x, t_{\max}) = 0$を初期条件に、時間$t_{\max}$から過去に向かって時間0まで数値計算することによって、$\lambda_D(0)$と$\lambda_c(x, 0)$が得られるわけである。$\lambda_c(x, 0)$と$\lambda_D(0)$は、それぞれt=0におけるIの勾配：$\partial I/\partial c(x, 0)$と$\partial I/\partial D(0)$であり、$\partial I/\partial c(x, 0)$と$\partial I/\partial D(0)$が得られたということは、勾配法によって、$c(x, 0)$と$D(0)$を数値計算にて算出できることを意味する（つまり初期濃度場推定およびパラメータ推定が可能となった）。ここで、数値計算手順についてまとめてみよう。

4.2 アジョイント法における数値計算の流れ

アジョイント法における一連の数値計算の流れは、具体的に以下のようにまとめられる。

(a) 濃度プロファイルの実験データの入手

通常の拡散対実験等から、例えば、5つの時間における濃度プロファイル：$c_{\text{obs.}}(x, t_1)$ ～$c_{\text{obs.}}(x, t_5)$が得られているとしよう。また初期濃度プロファイルの真値を$c_{\text{true}}(x, 0)$、また拡散係数の真値を$D_{\text{true}}(0)$とする。したがって、この場合の問題設定としては、$c_{\text{obs.}}(x, t_1)$～$c_{\text{obs.}}(x, t_5)$のデータを用いて、$c_{\text{true}}(x, 0)$と$D_{\text{true}}(0)$の値を推定する問題となる（なお通常の拡散対実験では、$c_{\text{true}}(x, 0)$は既知であるので、$D_{\text{true}}(0)$のみを推定する問題となる）。

(b) $c_{\text{true}}(x, 0)$と異なる初期条件$c(x, 0)$、および$D_{\text{true}}(0)$と異なる拡散係数$D(0)$を用いて、$c(x, t)$を計算する（通常の拡散方程式に基づくシミュレーションを行う）。

(c) この問題の場合、$(\partial J/\partial D) = 0$であることを考慮して、アジョイント方程式は、式(4)より、

$$\frac{\partial \lambda_c}{\partial t} = -D\left(\frac{\partial^2 \lambda_c}{\partial x^2}\right) - \frac{\partial J}{\partial c} = -D\left(\frac{\partial^2 \lambda_c}{\partial x^2}\right) - w(t)\{c(x,t) - c_{\text{obs.}}(x,t)\},$$
$$\frac{\partial \lambda_D}{\partial t} = -\int_x \lambda_c\left(\frac{\partial^2 c}{\partial x^2}\right)dx,$$

となる。また未定乗数の初期条件は、式(4)で与えられている$\lambda_c(x, t_{\max}) = 0$および$\lambda_D(t_{\max}) = 0$である。$c(x, t)$には、(b)で計算された値を用い、$c_{\text{obs.}}(x, t)$には、(a)の$c_{\text{obs.}}(x, t_1)$～$c_{\text{obs.}}(x, t_5)$を用いる。アジョイント方程式を$t_{\max}$から過去に向かって時間0まで数値計算することによって、$\lambda_D(0)$と$\lambda_c(x, 0)$が求まる。$\lambda_c(x, 0) = \partial I/\partial c(x, 0)$および$\lambda_D(0) = \partial I/\partial D(0)$であるので、勾配法を用いて、例えば、

$$c(x,0) \Leftarrow c(x,0) - \varepsilon\frac{\partial I}{\partial c(x,0)} = c(x,0) - \varepsilon\lambda_c(x,0),$$
$$D(0) \Leftarrow D(0) - \varepsilon\frac{\partial I}{\partial D(0)} = D(0) - \varepsilon\lambda_D(0)$$

にて、(b)で初期設定した$c(x, 0)$と$D(0)$を順次修正する（ε

は正の定数)。以上をc (x, 0) とD (0) が収束するまで繰り返し計算することによって、c_{true} (x, 0) およびD_{true} (0) が得られる。なお最後の部分の繰り返し計算には、単純な勾配法以外に、準ニュートン法などの各種の収束アルゴリズムを活用することができる。

4.3 アジョイント法の長所と短所

以上がアジョイント法の流れであるが、この理論は、変分法により停留値問題を解いているわけではない点に注意されたい。変分は、あくまで、関係式と条件 (アジョイント方程式と、アジョイント方程式を解くための初期条件) を導くために利用されている。アジョイント法の利点は、探索される解空間が、もともとの微分方程式 (この場合は拡散方程式) にて限定される点にある。つまり無駄な計算が少なく、大自由度系を対象としたパラメータ推定問題に適している。一方、この手法の欠点は、対象とする方程式系が変われば、アジョイント方程式自体も変わるために、モデルごとに理論式から全てを構築しなおさなくてはならない点である。つまり汎用プログラムを作ることが極めて難しい。

ここで少し、フェーズフィールド法[20,21] とアジョイント法との関係[22] に触れておこう。上述のアジョイント法の欠点は広く知られているアジョイント法最大の短所であるが、実は、アジョイント法のフェーズフィールド法への適用を考えた場合、この部分は欠点にならない。もともとのフェーズフィールド法自体が、現象ごとに方程式そのものから定式化しなおす場合がほとんどであるので、フェーズフィールド法へのアジョイント法の適用を想定した場合、プログラミングの手間はふつうのフェーズフィールド法を構築する場合と同等である。また優れたフェーズフィールドモデルほど、解空間がしっかりしているので、優れたアジョイント方程式に生まれ変わる。以上では、拡散係数と初期濃度場の推定を例に定式化を説明したが、フェーズフィールド法には、拡散係数だけでなく、材料工学における様々な定数が内在されている。これら全てが同化対象となるので、本手法の有用性は極めて高い。

フェーズフィールド法へのアジョイント法の適用について言及したが、アジョイント法は、微分方程式にて現象を記述する問題に普遍的に適用できる数学的手法である。またデータ同化プロセスの基本変数として、上記の例では時間tの場合を取り上げたが、微分方程式の形に記述できていれば、時間変数にこだわる必要も無い。時間に関する微分方程式だけでなく、温度に関する微分方程式や、ひずみに関する微分方程式など、材料学には、多くの応用先があると思われる。

5 材料工学とデータサイエンス

最後にデータサイエンスの手法を、材料工学へ適用する場合のキーポイントについて、やや私見もふくめてまとめてみよう。

5.1 逆問題によるパラメータ推定

材料工学では、各種の構成式や理論・シミュレーション内に、値が明確でない材料パラメータが存在する場合が多い。機械学習[14-16] は逆問題に関する数学的手法の宝庫であるので、逆問題による材料パラメータ値の推定は有効である。特に近年の計測分析装置の革新とシミュレーション手法の高度化は、逆問題解析の可能性を大きく後押ししている。シミュレーション結果と実験結果が一致する確率を用いて、計算内で用いた各種定数値の尤度 (著者は、"尤度"は"信頼度"と記した方がわかりやすいと思う) をベイズ推定する手法は、材料工学に多大な恩恵をもたらすと思われる。

5.2 特性に関わる各種因子の感度解析

材料開発の中心課題の一つは特性の最適化であるが、材料開発では、特性に影響する諸因子が多種多様である場合が多く、また条件によってその重要度が入れ替わる場合も珍しくない。したがって、特性に影響する因子の優先順位と個々の因子の特性に及ぼす寄与率 (感度) を、迅速かつ定量的に知る方法論の確立は、材料開発の根本的課題である。さて上記の逆解析は、結果と因子とを結びつける方法論であるので、例えば、パラメータ値の変化と、特性の変動の相関も同時に解析できる。当然ながら、パラメータ値の微小変化で特性が大変動する場合、そのパラメータは主要な因子であろう。つまり上記の逆解析手法は諸因子の感度解析を内包している。

5.3 仮想スクリーニング

仮想スクリーニング[5,23] は、計算精度が高いが計算負荷が大きいシミュレーションが存在する分野で威力を発揮する。手順は以下のとおりであり、有用性は明らかだろう。(1) 代表的なパラメータセットについて、通常のシミュレーションを実行し、入力と出力のデータセットを作成する。(2) これらデータセットを機械学習させ、シミュレーションの簡易コピーを作成する (たとえば、ニューラルネットに学習させたとしよう)。(3) 学習済ニューラルネットにて広域探索し、候補や条件を選出する。(4) 得られた候補や条件に対して、再度、元のシミュレーションで結果を確認する。なおシミュレーションの簡易システム自体が得られる点も有益であろう。

5.4 データの認識と分類

これは最適化対象が複雑で、その分類が困難である場合に起こりがちな課題ある。鉄鋼組織などはその典型で、例えばパーライト、ベイナイト、マルテンサイトといっても、多様な形態が存在し、現在においても、これら組織を完全に識別する数学的指標は無い（通常、光学顕微鏡等で、目で見て判断している）。材料組織の最適化を実現する場合、対象が明確に定義されなければ、精緻な最適化が不可能なことは自明である。問題の根源は、複雑・多様な材料組織の画像認識の問題であるが、この問題は、機械学習における画像認識技術の発展[24]によってほぼ解消された。今後、相変態・析出組織、再結晶組織、変形組織、結晶方位組織等々、適用対象は無限であろう。

5.5 検量線の非線形内挿

あえて検量線と記したが、実験的に決定されてきた各種の回帰関数の非線形精緻化の問題である。深層学習[14-16]により、多変数・非線形なデータの高精度回帰、近似が甘い部分の自動同定（ガウシアン・プロセス）、および超空間における領域分離などを容易に実現できるようになった。たとえば、Ms点、キュリー点、降伏点、および破壊靭性などを、1000くらいのパラメータ空間で、容易かつ精緻に非線形近似できたら、鉄鋼材料のスクリーニングに革命が起こるかもしれない。なお通常の多変量解析などによっても同様の試みは不可能ではないと思われるが、近年の機械学習が、無限開放型データを対象とした学習である点が、決定的な差となる点を強調しておきたい。つまり再学習が容易なのである。たとえば、ある3成分系のMs点を組成の関数としてニューラルネット近似し、これを4成分系に拡張する場合、3成分系の学習済みニューラルネットから学習をスタートすれば、完全に積み上げ方式で、効率よく学習を進めることができる。このように、既存の学習済みニューラルネットを踏み台とする効率的ニューラルネット学習法は、転移学習[16]として知られている（成分系をパラメータとした学習の全てに適用できる概念であるので、材料学（特に鉄鋼材料）での有用性は計り知れないだろう）。

5.6 重要データの識別

ビッグデータそのものを自在に操るノウハウは重要であるが、重要なデータと無意味なデータを選別する手法はさらに重要である。この手法としてはスパース学習[16,25]が典型例であろう。ラッソ解析はペナルティー項のパラメータ空間における軸上を効果的に活用して不要データをそぎ落とす有益な手法であるが、材料工学分野のスパース学習で最も重要な部分は、コスト関数（名称が分野によって異なるが、たとえば二乗誤差を計算する部分が対応）内の"材料モデル"である。理由は単純で、神技のような「材料モデル」があればデータは不要である（すべて計算で出してしまえばよい、究極のスパースモデルである）。したがって、材料工学の観点からは、当面、注力すべきはまず材料モデル側であり、それを補助する役割が、スパース学習である点を強調しておきたい。

5.7 組み合わせ爆発への対応

合金の世界では、成分数の増加による計算量の急拡大問題、いわゆる組み合わせ爆発の問題がいたるところで発生する。先に記した仮想スクリーニングと方法論は同じであるが、近年、ニューラルネット学習の精度が格段に向上したので、たとえば、多成分系のギブスエネルギーや化学ポテンシャルを、ニューラルネット近似して活用する試みが始まっている[26]。1点の組成にかかる計算は短時間でも、相分離シミュレーションのように、組織内の全ての点で、ギブスエネルギーや化学ポテンシャルを算出する場合、特に多成分系では計算量は莫大となる。この部分を計算負荷の軽い学習済みニューラルネットに置き換えることによって、この問題は解決されることが明らかとなってきた。また、ここで量産される学習済みニューラルネットは、多成分系に対するギブスエネルギーや化学ポテンシャルの簡易高速計算システムそのものであるので、それ自体が価値を生む可能性も高い。もちろん、成分系を増やす際の転移学習まで考慮すると、その価値は拡大する一方である。

5.8 不要なデータや間接データの重要性

データに関するパラダイムシフトは、これまで不要と思われていたデータも意味を持つことである。材料開発の実験で思わしい結果が出なかった場合、通常そのデータは廃棄される。しかし機械学習を前提とする場合、最適解近傍のデータはもちろん必要であるが、最適解から離れたデータも必要である（これが無いと最適解であることが機械学習できない）。さらに間接的な実験情報（実験装置自体の温度や試験時の音など）も、プロセス最適化の観点からは、データ同化の対象になり得る。つまり、これまで全く活用されてこなかった情報が、探索・最適化において貴重なデータとなる可能性が浮上している。IoTの発展も含め、材料設計・プロセス設計の方向性に本質的変革が生まれる可能性が高い。例えば、鉄鋼の圧延ラインの至る所の温度をサーモグラフで連続撮影すると同時に、各所の音も連続記録し、これらのビッグデータと、鉄鋼製品の各種特徴量を深層学習させれば、製造プロセス最適化に関する種々のヒントが得られるように思う。

5.9 言語処理と材料工学

言語処理における機械学習の構造は、実は材料工学における機械学習の構造に親和性が高いと思われる。通常、言語処理では、文字、単語、文、パラグラフを階層構造として認識し、

書かれている内容・意味を判断していく[27]。一方、材料工学では、原子、単位胞、相、組織、部材を階層構造として認識し、材料の特性・用途を判断している。このように、材料工学と類似な枠組みが、機械学習の様々な分野に認められる場合があり、今後大きな展開があるかもしれない。

6 おわりに

最近、MIに関して多くの事例が報告されてきているが[5,28]、材料技術者の目で全体を俯瞰した視点が少ないように思い、本稿では、著者が重要と考える点について、雑駁ではあるが説明させていただいた。材料工学において、やはり将来的にも試行錯誤的開発手法が消えることは無いと思う。しかし試行錯誤の効率を向上させることは可能であり、データサイエンスの機械学習は、この効率を圧倒的に高める新しいツール群である。材料開発の基本的アプローチが新しいステージに入った感があり、また材料の中でも群を抜いて複雑な相変態を有する鉄鋼材料は、当該分野の恩恵を最も享受できる材料と、率直に思うのだが、いかがであろうか。

謝辞

本稿執筆において、JST SIP (マテリアルズインテグレーションシステムの開発)、JST PRESTO (課題番号JPMJPR15NB)、新学術領域研究ハイエントロピー合金(課題番号18H05454)、元素戦略磁性材料研究拠点、MI^{2}I (情報統合型物質・材料研究イニシアティブ) 等にて、折にふれて議論させていただきました内容を参考としました。

参考文献

1) 研究開発の俯瞰報告書ナノテクノロジー・材料分野, (2017), http://www.jst.go.jp/crds/pdf/2016/FR/CRDS-FY2016-FR-05.pdf, (2017年5月現在)
2) Modeling Across Scales：A Roadmapping Study for Connecting Materials Models and Simulations Across Length and Time Scales, www.tms.org/multiscalestudy, (2017年5月現在)
3) K.Rajan (編著)：Informatics for Materials Science and Engineering, Elsevier, (2013)
4) Microstructure Informatics in Process–Structure–Property Relations, MRS Bulletin, 41 (2016)
5) Nanoinformatics, ed. by I.Tanaka, Springer, (2018)
6) 足立吉隆 (主査)：鉄鋼インフォマティクス研究会成果報告書, 日本鉄鋼協会, (2017)
7) Wikipedia, https://en.wikipedia.org/wiki/informatics, (2018年8月現在)
8) Wikipedia, https://en.wikipedia.org/wiki/bioinformatics, https://en.wikipedia.org/wiki/Bioinformatics, (2018年8月現在)
9) Wikipedia, https://en.wikipedia.org/wiki/Big_data, (2018年8月現在)
10) 樋口知之：予測にいかす統計モデリングの基本―ベイズ統計入門から応用まで, 講談社, (2011)
11) 花田憲三：実務にすぐ役立つ実践的多変量解析法, 日科技連出版社, (2006)
12) 花田憲三：実務にすぐ役立つ実践的実験計画法, 日科技連出版社, (2004)
13) S.B.McGrayne (原著), 冨永星 (訳)：異端の統計学 ベイズ, 草思社, (2013)
14) 杉山 将：イラストで学ぶ機械学習, 講談社, (2013)
15) C.M.ビショップ (著), 元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇 (監訳)：パターン認識と機械学習, 丸善, (2012)
16) たとえば, 機械学習プロフェッショナルシリーズ, 講談社.
17) 樋口知之 (編著)：データ同化入門 (予測と発見の科学), 朝倉書店, (2011)
18) 淡路敏之, 池田元美, 石川洋一, 蒲地政文：データ同化―観測・実験とモデルを融合するイノベーション, 京都大学学術出版会, (2009)
19) 久保拓弥：データ解析のための統計モデリング入門, 岩波書店, (2012)
20) 小山敏幸, 高木知弘:フェーズフィールド法入門, 丸善, (2013)
21) 足立吉隆, 小山敏幸：3D材料組織・特性解析の基礎と応用, 新家光雄 (編), 内田老鶴圃, (2014)
22) S.Ito, H.Nagao, A.Yamanaka, Y.Tsukada, T.Koyama, M.Kano and J.Inoue：Physical Review E, 94, 043307, (2016)
23) A.Seko, A.Togo, H.Hayashi, K.Tsuda, L.Chaput and I.Tanaka：Phys. Rev. Lett. 115 (2015), 205901.
24) 足立吉隆, 松下康弘, 上村逸郎, 井上純哉：システム/制御/情報, 61 (2017), 188.
25) 冨岡亮太：スパース性に基づく機械学習 (機械学習プロフェッショナルシリーズ), 講談社, (2015)
26) 野本祐春, 瀬川正仁, 若目田寛, 小山敏幸, 山中晃徳：日本計算工学会 第23回計算工学講演会論文集, B-04-04, (2018)
27) 高村大也 (著), 奥村学 (監修)：言語処理のための機械学習入門, コロナ社, (2010)
28) マテリアルズ・インフォマティクス～データ科学と計算・実験の融合による材料開発～, 情報機構, (2018)

(2018年8月31日受付)