総説

DeepInsight法:ゲノミクスなど非画像データを深層学習で扱う方法

角田達彦 東京大学大学院理学系研究科生物科学専攻医科学数理研究室

Although one would expect to use genomic data to discriminate diseases/phenotypes, this is a complex and hard task as key information to achieve it is spread over high-dimensional spaces. To address this issue, we envisaged that clustering similar variables, e.g. genes, would facilitate the identification of such information. Thus, here we have developed DeepInsight, a method comprising three steps: rearrange variables, extract features, and construct a classification model. Non-image data are transformed into image data allowing us to apply and take advantage of the learning capabilities of Convolutional Neural Networks (CNNs). Our experiments showed that with real data, such as cancer transcriptome, DeepInsight achieved much higher accuracy than typical machine learning methods. Our methodology has the potential to contribute to various biomedical applications and improve feature and structure extraction in non-image data.

Deep learning / convolutional neural network / genomics / feature extraction / classification / cancer

1. はじめに

ゲノムなどのデータは、病気や治療法の個人差などの解析・診断に役立つと考えられている。しかしそれらのデータは、数万から数千万の変数を持つ超高次元データであることが多く、伝統的な統計学では解析が難しい課題が多く存在する。これを克服するために、近年では「機械学習」に期待が集まっている。機械学習の一つの応用例として、機械が多くの変数の中から重要と判断した変数群を抽出し、その変数群を用いて、新規のサンプルが特定の病気と関連があるか否かを判別することが挙げられる。しかし、ゲノムなどのデータには背後に複雑な構造があることから、典型的な機械学習では簡単に取り扱えない。機械学習の多くでは、そのような複雑な構造はあらかじめ解いて、互いに独立な因子にした（特徴抽出）後の段階から使うと、精度が高くなる場合が多い。すなわち、前段階の特徴抽出をいかにうまく行うかがポイントになる。

そこで筆者らは、その特徴抽出を自動的に行えないかと考え、機械学習の中でも、現在の人工知能技術の主流の一つであり、特徴抽出能力を自前で持つ「深層学習」に着目した。深層学習を行うモデルの一つである「畳み込みニューラルネットワーク (CNN)」では、各サンプルは画像の形で入力し、隠れ層によって特徴抽出と分類を行う。このモデルの利点は、第一に、データから自動的に特徴を導き出すため、追加の特徴

抽出手法は必要ない。第二に、画像の高次の統計量や非線形相関などの複雑な構造が見つけれられる。第三に、畳み込みニューロンはその受容野などの限られた部分空間のデータを処理するので、大きな入力サイズに対して非常に多数のニューロンを持つ必要がなくなり、より少ないパラメータでも、ネットワークをより深くすることができる。

さて、実際の空間では多くの場合、近接する部分同士には何らかの類似性がある。同様に、写真や絵などの画像にも、互いに近接する画素は類似の情報を持つ傾向にある、という秩序がある。CNNでは、画素ごとに処理を行う際に周囲の情報も使って処理するため、画像内の構造を捉え、特徴を抽出することが可能になる。だが画素の配置が不適切だと、CNNの特徴抽出とその後の分類性能に悪影響が出る可能性がある。ここで、このCNNによりゲノムなどのデータを解析するには課題が生じる。それらのデータは非画像形式であり、隣り合う変数同士では明確な関係性が見られない場合が多い。CNNは入力として画像データを必要とするため、それらを直接は使用できない。ただし、データ同士の関連を適切に考慮しつつ非画像データを画像データに変換できれば、CNNを使って特徴抽出と学習ができ、典型的な機械学習よりも分類性能を向上させることが期待される。そのためには、画素としての変数の配置をうまく行う方法を考える必要がある。筆者らは、類似の要素や機能をまとめて配

DeepInsight: Methodology to Handle Non-image Data, such as Genomics Data, with Deep Learning

Tatsuhiko TSUNODA

Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo

置し、異なるものをさらに離して配置することによって、隣接する要素をまとめて使えるようにした画像を作成し、CNNを利用して特徴の抽出と分類を行うという3つのステップ（要素配置、特徴抽出、分類）を統合したDeepInsight法¹⁾を提案した。その詳細を以下に示す。

2. DeepInsight 法—アルゴリズム

筆者らは、データを使って診断や予測のためのクラス同定・分類を高精度に行うためには、「適切な変数（画素）の配置」、「特徴抽出」、そして「適切な分類モデルの構築」、という3つのステップを経る必要があると考え、それらを統合したDeepInsight法を提案した。

まず、変数の配置に関して、近傍情報を無視して独立に変数を扱うよりも、類似した変数や生データをクラスタとして変換する方が、周囲の重要な情報を互いに補完・統合できるため、信頼性が高くなる。そこで、DeepInsight法では最初に、似た変数をまとめて配置し、異なるものを離して配置することによって、隣接する変数をまとめて画素集合のように使えるようにした画像を作成する。例として、図1aに遺伝子発現値からなる変数ベクトル x を変換 T で変数行列 M に変換する全体像を示す。変数ベクトルの変数の位置は、変数の類似性によって決める。例えば、計算の結果類似性が高い変数 g_1, g_3, g_6, g_d 同士が互いに近くなるように、各変数の位置をマトリクス内で決める。そしてそれらの変数の値を各位置に代入する。これにより、サンプルごとに一つずつ画像が生成される。つまり、 d 次元変数を持つ N 個のサンプルが、 $m \times n$ 行列（ $m \cdot n \geq d$ ）を持つ N 個のサンプルに変換される。図1bに、より具体的な手順を示す。

学習セットは、各変数の位置（点）を決めるために用いる。 n 個のサンプルからなる学習セットを $X = \{x_1, x_2, \dots, x_n\}$ 、もとの変数ベクトル x_i は d 個の変数を持つとし、遺伝子または変数セット $G = \{g_1, g_2, \dots, g_d\}$ を定義する。ここで、 $g \in R^d$ 、すなわち、特徴 g_j は、 n 個の学習サンプルを有する。 G は X を転置することによって得ることができる。この変数集合 G に対し、t-SNE²⁾やカーネル主成分分析（kPCA）³⁾などの類似度計算手法や次元圧縮手法を用い2次元平面を得ることで、各変数の位置（点）を決める。ただし通常次元低減を行うと何かの情報を失うが、本手法では、2次元平面内で変数の位置が重ならなければ情報は欠落しない。このデカルト平面上の点は変換後の変数、または遺伝子に対応する。これらの点は変数の位置を定義するだけで、変数値や表現値を定義するのではない。

変数の位置が定義されたら、凸包アルゴリズムを使

図 1

変数ベクトル x を変換 T で行列に変換する全体像と変換の具体的な手順。a) 遺伝子発現値からなる変数ベクトル x を、変換 T で変数行列 M に変換する。b) 変数ベクトルを画像の 픽셀に変換する方法。CNNに入力するために、凸包アルゴリズムで全ての点を含む最小の長方形を見つけ、回転させ水平・垂直の形式にし、各座標を 픽셀にマップする。文献1を改変。

用して全ての点を含む最小の長方形を見つける。CNNアーキテクチャの場合、画像は水平または垂直の形式で囲む必要があるため、回転をさせる。そしてデカルト座標を 픽셀に変換する。このとき、画像サイズに 픽셀制限があるため、いくつかの変数を平均する。したがって、 픽셀フレームは、サンプル x_j ($j = 1, 2, \dots, n$) の変数（または遺伝子）の位置で構成される。

位置が決定したので、変数（または遺伝子発現）の実際の値をこれらの 픽셀位置にマッピングする。複数の変数（遺伝子）が 픽셀フレーム内の同じ位置に対応する場合には、それぞれの変数を平均化し、同じ位置に配置する。このように変数ベクトルを画像に変換すると、CNNで特徴抽出を行い、分類や予測をすることができるようになる。

図2aに、DeepInsight法により、2種類のがん検体を変換した画像の例を示す。

3. CNN アーキテクチャ

特徴ベクトルが画像に変換されれば（図2aなど）、CNNアーキテクチャでさらに処理することができる。今回、CNNのアーキテクチャとして、効率的にモデ

図 2

DeepInsight 法で変換したサンプルの像と使用した深層学習のアーキテクチャ。a) 2 種類のがん検体を、図 1 で示した DeepInsight 法で変換し、その画像を深層学習に入れる。b) 設計した CNN アーキテクチャ。画像を入力データとし、分類結果を出力するように学習する。効率的にモデルを学習させるため、異なるサイズのフィルタを用いた処理を同時に行えるように、上下 2 段の構成にした。文献 1 を改変。

るを学習させるため、異なるフィルタサイズを同時に扱えるものを設計した (図 2b)。このアーキテクチャでは、2D 畳み込み、バッチ正規化、ReLU アクティベーション、最大プーリングからなる多層を 1 パートとし、それが 4 つずつ、上下にある。バッチ正規化は学習中の過剰適合を防ぐために用い、最大プーリングレイヤは各レイヤの画像サイズをダウンサンプリングするために用いる。最後に、SoftMax を用いてクラス分類結果を出力する。

ここで、フィルタの大きさなどのさまざまなハイパーパラメータがあり、それらを調整する必要があるが、それにはベイズ最適化技術を用い、検証セットで最適になる値を求めた。

図 3 に実際のデータを変換した例として、異なる種類のがん検体の遺伝子発現データ、異なる種類のテキストデータ、異なる種類の母音データを画像に変換したものを示す。それぞれのデータセットで左右を比較すると、違いが画像の上でわかるようになる。CNN の特徴抽出機能により、これらの違いが捕捉できていると考えられる。このようにして学習を終えると、新しいサンプルに対してもクラスを決定できる。

図 3

図 1 の方法で変換したサンプルの画像の実例。上段はがん、中段はテキスト、下段は母音のそれぞれの非画像データを画像データに変換した。文献 1 を改変。

4. DeepInsight 法の検証

次に、DeepInsight 法の検証を行った。そのため、「遺伝子発現データ」、「テキストデータ」、「母音データ」、「2 つの人工データ」のセットを用いて、既存の最先端の分類手法である random forest 法などと結果を比較した。「遺伝子発現データ」は、TCGA (<https://cancergenome.nih.gov>) からの公共データセットであり、10 種類のがんに対応する RNA-seq 遺伝子発現データの 6,216 のサンプルで、各サンプルには 60,483 の遺伝子発現値 (変数) が付けられている。「テキストデータ」には、オンラインのニュースグループの文書から抽出した Relathe というデータセットを用いた。これには 20 のニュースグループがあり、それぞれには 4,322 次元 (単語) の 1,427 の記事が入っている。ある記事の元のニュースグループを特定できるかどうかという 2 クラス分類問題を解くことになる。「母音データ」は TIMIT コーパスから得られた音声データセットである。音声は 10 個の異なる母音に割り振られ、次に各母音が 3 つのセグメントに細分されている。各セグメントはさらに特徴量に変換され、39 次元のデータになっている。これが合計 12,579 のサンプルに対して得られている。残りの 2 つは人工的なデータで、1 つめは 500 次元の連続値変数を持つ 2,600 個のサンプルが入っているデータセットである。2 クラス分類問題を解くためのコンテストの課題として作られたもので、多変量で高度非線形という特徴を持つ。そしてもう 1 つは Leo Breiman が考案したベンチマークデータであり、20 次元の 7,400 サンプルが含まれる。異なる平均値と共分散を持つ 2 クラスからなる。各サンプルの元のクラスを同定する課題である。

比較のため、random forestなどの既存の分類手法を用いた。各手法のハイパーパラメータは、グリッド検索最適化で最適化した。

これらを用いて検証した結果、遺伝子発現 (RNA-Seq) データのテストセットを用いたとき、random forest 法では分類精度 96% だったのに対し、DeepInsight 法では 99% を達成した。母音データでは、random forest 法の分類精度 90% に対して DeepInsight 法は 97%、テキストデータでは、random forest 法の分類精度 90% に対し、DeepInsight 法は 92% の精度を達成した。残りの 2 つの人工データセットでも同様の結果だった。さらに、5 つのデータセット全てを用いて平均分類精度を計算した結果、既存の技術のうち最高だった random forest 法の 86% に対し、DeepInsight 法は 95% の平均分類精度を記録し、はるかに優れたものであることがわかった。

5. おわりに

本研究は、ゲノミクス、母音、テキスト、人工的なデータセットなどのいろいろな非画像データを CNN で普遍的に扱えるようにした世界初の取り組みである。本手法を用いて、10 種類のがんのデータセットを用い、それらを互いに区別するように学習させた後で、分類する際にネットワークが重視した部分を可視化する手法である CAM⁴⁾ を使って、画像解析のように中間層の解釈を試みたところ、がんの特徴的な遺伝子群やパスウェイが多数抽出された。これは、当初の目的の一つである「特徴抽出」を CNN によって自動的に行うことが可能になったことを示している。すなわち、CNN がもともと得意とするクラス分類の学習のみならず、その特徴抽出機能によって科学的な発見も可能になるのではないかと考え研究を進めている。

また本手法により、非画像データの解析に対して、深層学習、特に CNN の、特徴抽出、次元削減、疎なデータや超高次元データからの隠れた構造の発見、少ないデータで学習するためのデータの水増しとアップサンプリング、ラベル付き/ラベルなしサンプルによる半教師あり学習、そして時系列データ解析などといった能力⁵⁾ を活用することができることがわかった。

さらに、DNA 配列やタンパク質配列、RNA-Seq などのさまざまなオミクスデータを深層学習によって解析できる可能性がある。それに加え、現時点で本手法は、入力が単層で 2 次元行列型の CNN を想定しているが、これを複数の層を組み込むように拡張すれば、例えば、遺伝子発現、メチル化、突然変異などのマル

チオミクスデータを扱う問題にも適用可能であろう。筆者らは、このように深層学習がより利活用されることにより、医学・生命科学の複雑なデータが紐解かれ、将来の個人ごとの診断や予測への貢献が可能となることを期待している。

本研究は、科学技術振興機構 (JST) 戦略的創造研究推進事業 (CREST) 「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化 (研究総括: 北海道大学・田中 譲)」における研究課題「医学・医療における臨床・全ゲノム・オミックスのビッグデータの解析に基づく疾患の原因探索・亜病態分類とリスク予測 (研究代表者: 角田 達彦)」, 文部科学省新学術領域研究「代謝アダプテーションのトランスオミックス解析 (領域代表者: 東京大学・黒田 真也)」の「次世代ヒト全ゲノム・オミックスの解析方法論の開発と応用 (研究代表者: 東京大学・角田 達彦)」, 東京医科歯科大学難治疾患研究所国際共同研究拠点活動による支援を受けて行った。

ソースコード、サンプルデータ、ドキュメント類は、Web サイトで入手可能である: https://www.riken.jp/en/research/labs/ims/med_sci_math/ または <http://www.alok-ai-lab.com>.

文 献

- 1) Sharma, A. *et al.* (2019) Sci. Rep. **9**, 11399. DOI: 10.1038/s41598-019-47765-6.
- 2) Maaten, L. J. P. V. D., Hinton, G. (2008) Journal of Machine Learning Research **9**, 2579-2605.
- 3) Schölkopf, B. *et al.* (1998) Neural Computation **10**, 1299-1319. DOI: 10.1162/089976698300017467.
- 4) Zhou, B. *et al.* (2016) IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2921-2929. DOI: 10.1109/CVPR.2016.319.
- 5) Goodfellow, I. *et al.* (2016) Deep Learning, The MIT Press. England.

角田達彦

角田達彦 (つのだ たつひこ)

東京大学大学院理学系研究科生物科学専攻教授
1989 年東京大学理学部物理学科卒業, 1991 年同大学院理学系研究科物理学専攻修士課程修了, 1995 年同大学院工学系研究科博士課程修了, 京都大学助手, 1997 年東京大学医科学研究所 RA, 1998 年 同助手, 2000 年理化学研究所遺伝子多型研究センターチームリーダー, ゲノム医科学研究センターグループディレクター, 統合生命医科学研究センターグループディレクター, 2015 年東京医科歯科大学難治疾患研究所教授, 2019 年より現職. 医学博士・工学博士.
研究内容: 医科学数理, ゲノム医科学
連絡先: 〒113-0032 東京都文京区弥生 2-11-16 東京大学理学部 3 号館 407
E-mail: tsunoda@bs.s.u-tokyo.ac.jp