退職記念特集

【随想】

木と森からモデルの予測性と理解しやすさについて考える

井上　弥

コロナ禍でマスクをしていると，iPhone の Face ID が顔を認識してくれないという問題はあるが，ディープ・ラーニング（deep learning）技術のおかげで，顔などの認識率は格段によくなってきている。病院や施設によってはカメラの前に立つとマスクをしているかを識別し，体温まで測ってくれる装置が導入されていたりする。顔などの認識率が高いということは，学習されたモデルの予測精度が高くなっているということだ。高い予測率や計算コストの低下から，ディープ・ラーニングに代表されるような機械学習は様々な場面で使われるようになってきた。

木を見る

樹木モデル（tree based model）は，基準変数がカテゴリである場合の決定木（decision tree）や連続変量である場合の回帰木（regression tree）として知られているが，対応する（重）判別関数や重回帰とは異なり，直線性（線形性）を前提としないモデルとされる。その点で，適用できるデータの種類は多くなるだろう。

決定木を例にとると，所与のデータの説明変数ごとに，ジニ（Gini）係数などを用いて，より単純な基準変数のグループになるような説明変数を探して分割していき，最終的に基準変数のまとまったカテゴリに辿り着くように枝分かれした樹形図を作っていく。もちろん，1 グループ 1 ケースまで枝分かれすることも起こるので，分割されたグループのサイズなどの制約をつけたり，枝を刈り込んだりして樹形図を決定する。R（R Core Team, 2020）の樹木モデルである rpart（Therneau & Atkinson, 2019）では，minsplit，minbucket，cp（complexity parameter）で枝の成長（枝分かれ）を制御している。この樹形図は，重回帰などと同様に，各変数から最終的な分類を予測するモデルを構築したことになるのだが，重回帰などとは異なり，心理学の研究分野ではあまり見かけない気がする。

こうして出来あがったモデル（樹形図）から予測された分類と実際の基準変数の値は 100%一致するわけではない。一致の程度は，一致率（ヒット率）としてあらわされるが，これが高いほど，説明力のあるよいモデルといえるだろう。基準変数が連続変量である回帰木の場合，重回帰などと同じように，モデルの説明力として決定係数 R2 も出てくるが，決定木の場合は，この一致率が決定係数と対応するものだろう。ケース数にもよるようだが，よく使われる 5 段階評定も 7 段階評定と同じように間隔尺度（interval scale）として分析しても遜色ないようだから（萩生田・繁枡, 1996），決定係数を指標とすることの方が多いかもしれない。いずれにせよ，この一致率や決定係数が低いと，良いモデルとはいいがたいだろう。重回帰でも，個々の変数の偏回帰係数が有意で大きな値であろうと，モデルの説明力を表す決定係数 R2 が低ければ，モデルに投入した説明変数だけではほとんど説明できていないことになる。説明変数をみて，なかなか良いモデルができたと思っても，決定係数が小さいと，そもそもモデルとして成立していないかもとがっかりすることがある。

問題は，経験上たいていの場合，一致率や決定係数がそれほど高くないことにある。樹形図がわかりやすく，納得のいくものであったとしても，一致率や決定係数が低いということは，結局モデルでは説明できていないことになるので，あまり役に立たないモデルともいえる。マスクどころか，明るい部屋で，きっちり正面をみて，微笑んだり憮然としたりせずに普通顔で見つめないと認識してくれない感じだろうか。これは，実験ならいざ知らず，実生活ではほんとうに役に立たないだ

ろう。一致率や決定係数が高い方が，現象をうまく説明できる良いモデルと思える。

そうなると，いかに一致率や決定係数を高くして有用なモデルにするかに関心が向く。現実の世界では多種多様な要因が複雑に影響し合っているのだから，新たな説明変数を加えてモデルを再検討していくという手もある。しかし，そもそも研究を開始する時に，大きく影響する変数は入れ込んであるだろうから，そう簡単には新たな有力変数は見つからないだろう。頑張って，あまり影響しない変数をたくさん入れ込んでも，一致率や決定係数はそれ程大きくならないだろうし，説明変数の数を増やしすぎると，何がどう影響しているのかが，かえって判りづらくなるから，闇雲に増やしたくはない。

次に考えられるのは，樹形図そのものを改良する手である。つまり，もう少し予測精度の高いモデルを算出する方法を模索することである。先程の例でいえば，minsplit，minbucket，cp（complexity parameter）などの値を変えてみるというのもあるだろう。しかし，これも一致率や決定係数を大きく変えることはむずかしいだろう。

そこで登場するのが，ランダム・フォレスト（random forest）だ。これは，1 本の木では説明できなくとも，様々な木を組み合わせた森ならば，より説明力が上がるだろうという方法だ。もちろん，他の機械学習モデル，例えば R（R Core Team, 2020）でいえば，h2o（LeDell, Gill, Aiello, Fu, Candel, Click, & Kraljevic, 2020）のようなディープ・ラーニングでもいいのだが，隠れ層（hidden layer）をどうするべきかなどむずかしいことが多いので，ここでは樹木しばりということで，ランダム・フォレストにする。

森を見る

ランダム・フォレストは，データから複数の樹木を生成し，樹木の多数決で分類を決めるやり方と説明される。R（R Core Team, 2020）の randomForest（Liaw & Wiener, 2002）では，ntree で木の数（default では 500），すなわち森の大きさを指定し，mtry で木を作る時に用いる説明変数の数（その数だけランダム・サンプリングされるので，木の種類にでも当たるだろうか）を指定している。これによって，様々な木が ntree 本生成されることになる。そして，それらの木々が推定する分類を，森として多数決で決めるモデルといえるだろう。三人寄れば文殊の知恵というが，500 本も寄れば，菩薩を越えて如来の知恵にでもなるだろうか。

経験上，樹木モデルよりもランダム・フォレストの方が一致率や決定係数の値が大きくなる。つまり，よく説明できる役に立つモデルができあがる。ただし，モデルを構築するときに使ったデータの一致率を高め過ぎると，新たなデータでの一致率が落ちることがあるようなので，学習時の一致率の高さを無条件では喜べないようではあるが，木よりは森の方が予測性の高い良いモデルであろう。

機械学習では，モデルを作るためにデータを分析する過程を学習と呼ぶ。人ならぬコンピュータのプログラムが学習するから機械学習である。学習で得られた予測精度の高いモデルを新しいデータに当てはめて予測した時に，過剰に学習していると，予測の精度が落ちてしまうのが過剰学習だ。学習時の予測精度を高めようと過剰に学習すると，学習が細かすぎて，そのデータに特化したモデルになってしまうため，応用が利かなくなるのだろうか。機械学習なのに人間ぽくって，100%を目指して研究するよりも適度のいい加減さをもってた方がいいよと慰められているようで，ちょっと気持ちが温かくなる。

ところで，こうして得られた役に立つモデルを説明しようとすると厄介なことが起こる。樹木モデルが作った樹形図ならば，図に描くこともできたし（そもそも樹形だから），場合によっては，それを文章で記述することもできた。しかし，ランダム・フォレストが作った森を描くことはむずしい。それを文章で記述することも困難だ。確かに説明力のある良いモデルのはずなのだが，結局のところどういうモデルなのかを，簡単には説明できない。もちろん，構築されたモデルを新たなデータに適用することも可能なので，可搬性や実用性はあるのだが，簡単には説明できないモデルになる。重要そうな何本かの木を抜き出して描くことは可能なのだが，森そのものをうまく説明できない。つまり，良いモデルではあるだろうが，う

まく説明できないモデルであり，言ってしまえば，理解できないモデルなのである。

これを研究で使うと，どういう要因によって現象が起きるのかというメカニズムの説明ではなく，どういう説明変数を投入したら一致率や決定係数があがるのかや，どういう機械学習モデルを使うと一致率や決定係数があがるのかという記述になるだろう。何をどのように機械学習させるかが重要になる。しかし，こんなモデルでは，Pepper（1942）の分類ではメカニストにあたるだろう自分には，到底理解した気になれない。

木を見るべきか，森を見るべきか

わかりやすいけど精度の高くない木（tree）と，精度は高いけどわかりにくい森（forest），どちらが，われわれが求めているものなのだろうか。コンピュータが高速，高機能になり，低い計算コストで精度の高いモデルを構築できるようになったからこそ，立ち現れてきた問題のような気がする。学習開発学が目指す理論と実践の往還は，どちらのモデルを希求するのだろうか。理解しやすく説得力はあるけれどそれ程実用的でない木を目指すのか，わかりやすく説明はできないけれどよく当たる実用的な森を目指すのか。往還だから両方と言いたいところだが，両者はそう簡単に融合はできそうにない。いかに予測精度の良いモデル（森）を作るかが研究の目的になっている森の研究と，いかにわかりやすいモデル（木）を使って現象を研究するかを目的としている木の研究では，目指す方向が真逆のような気さえする。

木を見て森を見ず，森を見て木を見ずとはいうけれど，木も森も見られる研究者が求められているのだろうか。木と森を同時には見られない以上，機を見て，木を見たり，森を見たりしないといけないのだとも思う。自分の守備範囲を超えて，異なる視点から研究を見直すことが求められているのだろうし，そのためには新たなチャレンジが必要なのだろうが，ウロウロと落ち着かない気もする。

引用文献

萩生田 伸子・繁桝 算男（1996）. 順序付きカテゴリカルデータへの因子分析の適用に関するいくつかの注意点 心理学研究, 67, 1-8.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., & Malohlava, M. (2020). h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version 3.32.0.1. https://CRAN.R-project.org/package=h2o

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Pepper, S. C. (1942). World hypotheses. University of California.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Therneau, T. & Atkinson, B. (2019). rpart: Recursive partitioning and regression trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart