連載　統計的機械学習ことはじめ

廣野　元久　著

第２回　モデルの汎化性能の評価

今回は機械学習のモデル選択の際に使われる汎化性の評価について紹介します。

実験計画法では、計画段階で実験の普遍性を吟味し、実験後に必ず再現実験(確認実験とも言います)を行い、普遍性の評価を行います。それと同じような意味で、機械学習でもデータ分析プロセスの中で、普遍性の確認と同様な意味で汎化性の評価を行います。以下では、汎化性の評価の方法と必要性について考えてみたいと思います。

最初に、ソフトウェアの開発でよく使われるベリフィケーション(検証)とバリデーション(妥当性確認)の違いを説明します。データ分析のプロセスでは、入力情報はモデルを通じて予測や分類などの出力へ変換されます。入力に対して出力が正しく変換されているかを確認することをベリフィケーションと言います。これは主にモデルを開発する研究者の仕事です。

一方、出力がもともとの要求と合致しているかを確認することをバリデーションと言います。データ分析に対する要求とは、複数のモデル候補から学習データへのあてはまりの良さだけでなく、バイアスの影響にロバストな汎化性の高い最良のモデルを選択することです。このため、要求への合致度を評価するために、分析者はバリデーションを行う必要があります。

まとめると、出力の正しさを何に基づいて確認するかが両者で異なり、ベリフィケーションは入力、バリデーションは要求に基づいて、出力の正しさを確認するものです。図3はベリフィケーションとバリデーションの違いをイメージにしたものです。参考になさってください。

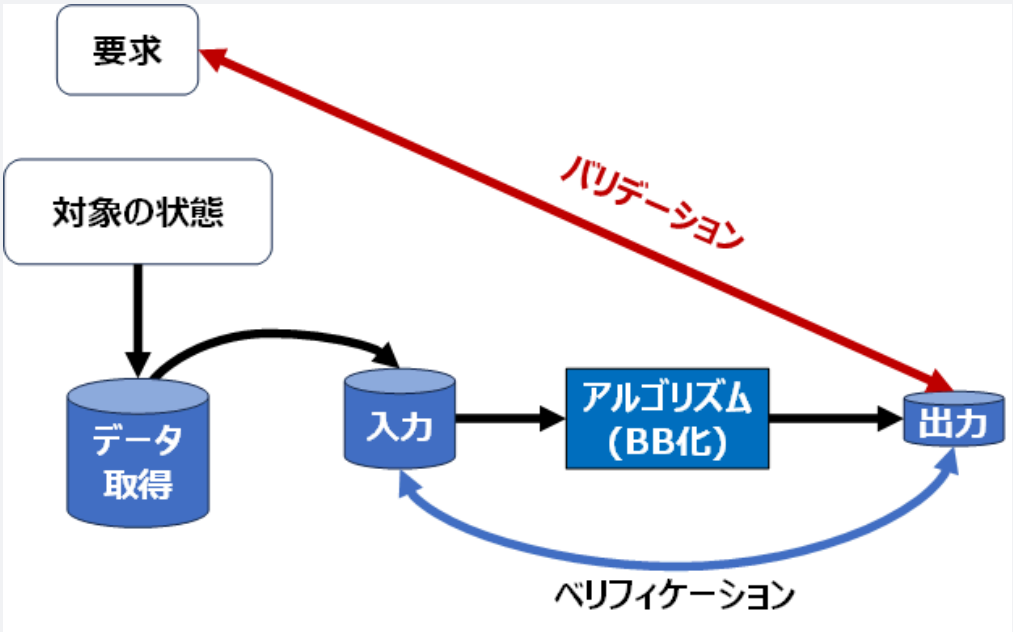

図３　ベリフィケーションとバリデーションの違い

（次ページへつづく）

近年、様々な様式のデータを活用できる環境が整ってきました。回帰分析でもビッグデータを扱うことが多くなりました。汎化性能を確認するには、個体数が十分に多いことを生かしてモデルの評価・選択を行います。クロスバリデーションは分析対象のデータを無作為に幾つかに分割し、モデルの推定・モデルの選択・モデルの汎化性をそれぞれ異なるデータで行います。ここでは、クロスバリデーションの中から最も基礎的なホールドアウト検証 ＊注) を紹介します。

図４は簡単なケースを使って、ホールドアウト検証の概念を示したものです。いま、全体の個体数 n＝300あるとします。簡単にするために全体を100個ずつ３グループに分割します。その中でモデルの推定に使うグループのデータを学習用データと言います。学習用データを使って、幾つかのモデルを推定します。このとき、与えられたデータに最もよくあてはまるようにモデルのパラメータを推定してしまうので、モデル探索時の推定バイアスが発生してしまいます。この推定バイアスに注意しないと、学習用データにはよくあてはまっていても、新しいデータを入力した場合に好ましい出力が得られない場合があります。

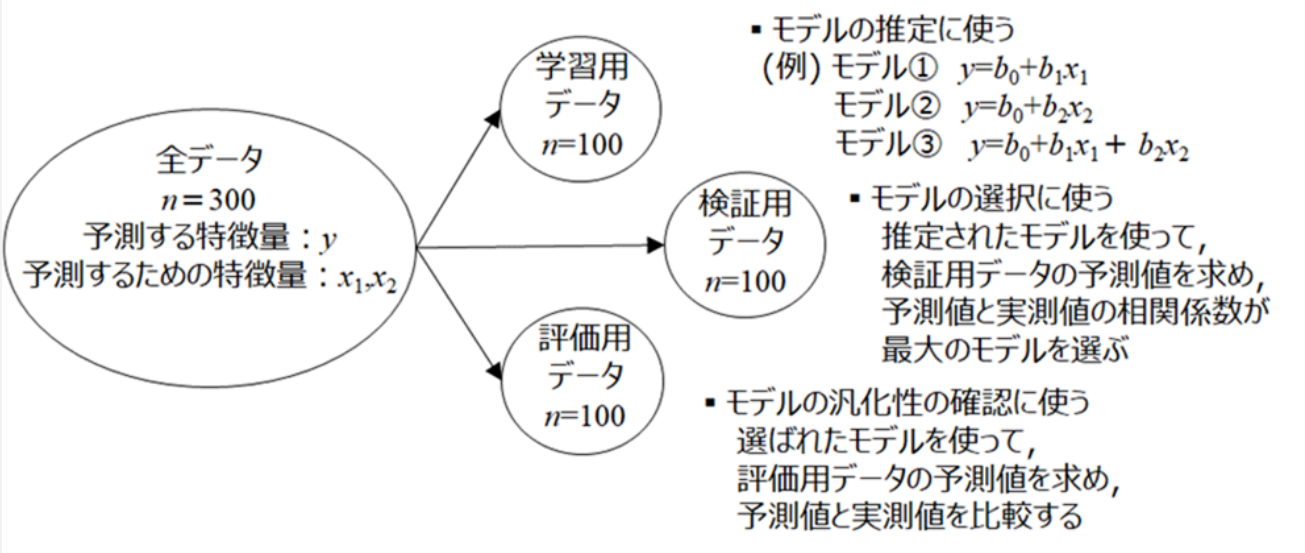

図４　ホールドアウト検証の概念図

そこで、推定バイアスの落とし穴にはまらないように別のグループを使ってモデルの選択を行います。このグループのデータを検証用データと言います。推定されたモデルの優劣を決めるために使うデータです。優劣の単純な決め方には、得られたモデルを使って検証用のデータの予測を行います。このとき、得られた予測値と実測値の相関係数の大きさでモデルの優劣を決めたりします。このプロセスでは、最適なモデルを選ぶ際に生じるモデル選択のバイアスに注意する必要があります。

そこで、最後のグループを使ってモデル選択のバイアスの影響を確認します。このグループを評価用データと言います。選ばれたモデルに評価用データを入力して予測や分類を行い、モデルがデータにどれだけあてはまっているかを調べます。そして、推定精度の目標を満足していれば、このモデルをアルゴリズムとして実装します。そうでない場合には元に戻って、新たなモデルの探索を行います。

（次ページへつづく）

では、なぜこのように面倒なモデルの性能を確認するステップが必要なのでしょうか。学習データで寄与率R^2や修正済寄与率 R^{*2} を使ってモデル検証をしておしまいでよいのではないでしょうか。多変量解析では扱うモデルは１つです。例えば、予測の問題では重回帰分析１つです。重回帰モデルの範囲内でモデルの改善は行うのですが、基本は一撃必殺です。しかし、機械学習では重回帰モデルだけでなく、SVM(サポートベクターマシン)、ニューラルネットワーク、ランダムフォレストなど、多数の様々なモデルを使って、予測や分類を試みます。あらゆる課題やデータに最高の精度を出すことのできる万能薬はありません。データ分析に対する要求やデータの性質により、それぞれのモデルの推定精度にはばらつきが生じます。モデルには得手不得手があるものです。要求やデータの性質が変わればモデルも変更すべきで、出来る限り先見知識を使って、その要求にあったモデルをアルゴリズムとして採用すべきです。ある特定のモデルだけをすべの要求に適用するのは極めて危険です。

このため、学習のステップではできるだけ多くの様々なモデルを用意して予選会を行います。予選会は結果のわかっているデータを使ってモデル構築やモデル選択をします。モデルはパラメータ数を増やして複雑な形にすれば、いくらでも推定精度は向上します。また、１つのモデルだけを選ぶのではなく、複数のモデルを採用して、予測値の平均を使うというアイデアも生まれています。これをアンサンブル効果と呼びます。予選会でのモデルは過学習になっている危険性、学習用データのみに過剰に適合し、意味のないサンプリング誤差に対して意味のある効果だと判断してモデルを複雑化し汎化性能が失われてしまう等の恐れがあります。このため、未知のデータに対するモデルの性能評価が必要になります。重複になりますが、推定精度の高い上位のモデルを少数選んで、検証用データでモデルの汎化性の評価を行い、検証用データでも高い予測精度が出たモデルを採用します。選ばれたモデルの評価をもう一度、別の評価用データで汎化性をさらに確認しているのです。

＊注)ホールドアウト検証

厳密に言えば、ホールドアウト検証はクロスバリデーションとして扱わないようです。

それは、バリデーションのステップでデータを交差(クロス)させることがないという理由からです。ご注意ください。

今回は読者の理解が進むように、ホールドアウト検証を汎化性能の評価法として紹介しました。

著者紹介

廣野　元久（ひろの　もとひさ）

1984年（株）リコー入社。以来、社内の品質マネジメント・信頼性管理の業務、SQCの啓蒙普及に従事、品質本部QM推進室長、NA事業部SF事業センター 所長を経て、現在、（株）リコー倫理審査委員会 委員。

東京理科大学工学部経営工学科 非常勤講師（1997～1998年）、慶應義塾大学総合政策学部　非常勤講師（2000～2004年）。（一財）日本科学技術連盟　多変量解析法運営委員会委員、講師。