連載　統計的機械学習ことはじめ

廣野　元久　著

第６回　ランダムフォレストの考え方

回帰分析や判別分析とは異なるアプローチとして、応答を逐次的に２分岐させて予測や分類を行う分岐モデル（ここでは木とよびます）を作る方法があります。その方法は決定木と呼ばれ、多変量解析の手法として古くから使われています。決定木はいつでも精度の高いモデルが得られるとは限りません。今回紹介するランダムフォレストは、その名が示す通り、様々な小さな木をランダムに発生させ、それぞれの木が示す結果を総合的に判断するアンサンブル学習を使った方法です。複数の小さな木を使って、強い森を作るのです。

最初に、出力（特性）が量的な場合の決定木を回帰モデルと比較しながら説明します。図14左はxとy の散布図です。yを予測するには方法①として回帰モデルがあります。図左では点線で回帰直線を、破線線で２次式をあてはめた結果が示されています。２次式をあてはめたほうがよいことがわかります。方法②が木モデルです。あるルールに従ってxの区間を設定して、区間ごとに平均を求めてステップ関数として繋ぐものです。今回はxの値３を境界として、３未満の区間の平均を求めて、その値9.1で平均線を引きます。同様に、３以上の区間の平均を求めて、その値13.1で平均線を引きます。その結果は細線のステップ関数で表すことができます。今度はxの値３以上の区間の中で、値４で２分割します。xの値が３以上４未満の区間の平均を計算して、その値9.3をこの区間の代表値と考え平均線を引き直します。このように順次xの値で境界を作り順次、ステップ関数として繋げていきます。その結果が図左の太線のステップ関数です。それを木で表したものが図右になります。実際の木とは似つかない奇妙な名前ですが、木とはデータ間の階層的な関係、つまり親と子のような関係を表現するのに用いられるデータ構造をさします。木は大量の情報を系統的に管理する際に有効です。木はノード（節）とブランチ（枝）から構成され、最上位のノードはルート（根）といいます。また、最下位の節をリーフ（葉）といいます。複数に枝分かれしている場合には、リーフは複数になります。ブランチはノードとノードをつなぐ経路になります。また、あるノードの上のブランチにつながるノードを親、下のブランチにつながるノードを子といいます。決定木で得られた結果は、全データを要素に分解（分類）することや、取り出す（予測）ことができるようになります。

（次ページへつづく）

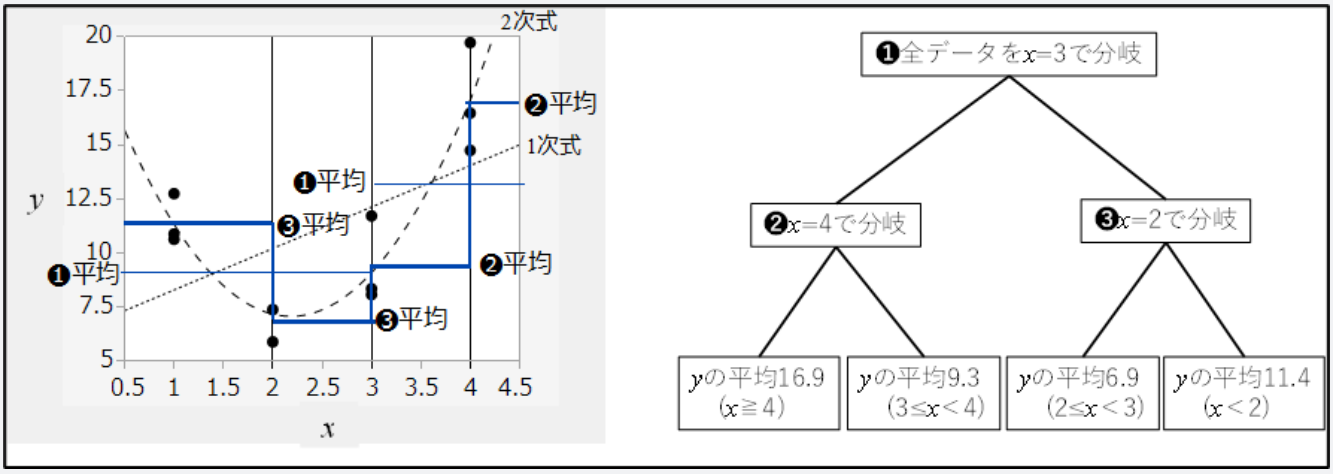

図14　決定木による予測例

決定木のイメージがつかめたところで、今度はランダムフォレストのお話です。ランダムフォレストは複数の木を組み合わせてあてはめを行う方法です。元のデータから復元無作為抽出した要素のデータで決定木を作るという処理を何度も行います。各木の各分岐では無作為抽出された指定個数の特徴量の中から分岐で使用される特徴量が選定されます。このようにして得られる多くの木を組み合わせて、最終的に予測精度の高い結果が得られるという考え方です。最終的な予測はすべての木から得られる予測値を平均したものになります。このような考え方をアンサンブル学習といいます。

話をわかりやすくするために、今度は出力が質的な場合（クラス C_1 と C_2 に分ける場合）を考えます。図15はランダムフォレストでの分類の考え方を示したものです。出力が質的な場合は、様々な決定木で得られた判定結果の多数決により個体の属するグループが決定されます。この場合は多数決で C_2 に属すると分類されます。

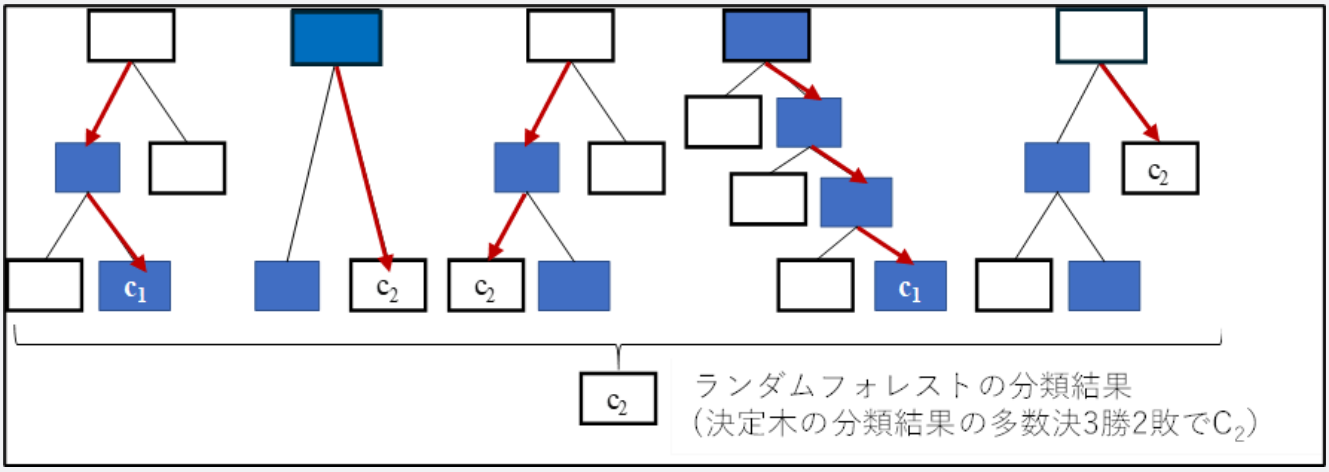

図15　出力が質的な場合の個体が属するクラスの決定方法

（次ページへつづく）

また、ランダムフォレストで設定が必要なハイパーパラメータの中には、値が大きいほど汎化能力が強くなる性質がありますが、大きくしすぎても汎化能力は衰えにくいため過学習のリスクが小さいことが知られています。また、汎化能力と特徴量の重要度が簡単に計算できるというメリットもあります。しかし、各個体は複数の木の結果の多数決により分類先が決まるので、決定木のような解釈を行うことができません。なお、ランダムフォレストのハイパーパラメータは、一般的に木の数・各木で分岐条件に使う特徴量の数・各木を作成する際に使う個体数・木の設定（剪定の閾値・木の深さ・ノードの最小個体数）などがあります。

以上、これまで統計的機械学習の一般的な手法の概要を６回に渡って説明いたしました。ご紹介できた手法に関しては、ご自身が持っているデータを使って、解析手順を学んでいただければ幸いです。機械学習の手法は日進月歩で進化しており、ご紹介しきれなかった手法も沢山あります。それらについては成書をお手に取り、解析手順を身につけていただければと思います。毎回お読みいただき、どうもありがとうございました。

（以下余白）

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。