応用統計学 Vol. 40, No. 3 (2011), 193–210

研究論文

Random Forestを用いた欠測データの補完に基づく大学入試センター試験科目間得点差

大学入試センター 石 岡 恒 憲

要 旨 Breiman によって提案された分類や非線形回帰のための集団学習の方法の一つである Random Forest（RF) が，欠測を多く含む大量データに対して安定してかつ精度のよいデータ補完 (imputation) を実施することを示す．本報告では，RF によるデータ補完の方法について解説し，ある年度のセンター試験の理科および社会の科目間難易比較についての応用例を示す．説明変数が全て同等もしくは同列ではなく，幾つかの説明変数がグループにまとめられ，またそのグループの中から一つが排他的に選択されるような場合には本報告の手順は有効であろう．

1. はじめに

平成 2 年度 (1990) より大学入試センター試験はアラカルト方式が採用され，それまでの 5 教科 7 科目の受験から任意の科目を必要なだけ受験できるようになった．これによりセンター試験は国公立大学受験者だけでなく私立大学受験者も受験対象となり，より広い学力層の人たちが受験するようになった．データ解析の観点からは，文科系志願者における「数学」や理科系志願者における「国語」など従来国公立大学受験者なら受験することが求められる教科において欠測データが生じることとなった．

またセンター試験は，教科内での科目選択の自由が認められており，たとえば「地理・歴史」教科では「世界史 B」「日本史 B」「地理 B」などの科目の中から受験時に 1 科目を選択する．選択しなかった科目はその受験者にとっては欠測データとなる．同様のことは「公民」教科の「現代社会」「倫理」「政治・経済」において，また「理科③」教科の「物理 I」「地学 I」においても存在する．

しかも欠測しているか否かの確率はランダムではない．通常，国公立大学を志願する者の方が私立大学専願者よりも一般的な学力は高いと考えられるから，私立文科系志願者における「数学」や私立理科系志願者における「国語」における欠測は他の科目の得点がより低いときに起きやすいと考えられる．以上のような

- 欠測割合の多さ
- より広い学力層の人が受験することによるデータを構成する母集団のモデリングの複雑さ
- 完全にランダムではない欠測メカニズム

に加えて，年間 50 万人という決して少なくない標本を取り扱うことの技術的な要請が我々にはある．

また現在行われている「地理・歴史」「公民」「理科」科目間の得点調整において，学力の高い層がたとえば「物理 I」を選択する傾向があるといった選択バイアスについての妥当的な検証についてもその要請がある．つまり現在しばしば問題となる科目間の得点調整において，科目間の平均得点差がその問題の難易度による差なのか，そもそもその科目を受ける受験者の所持する一般的な学力差によるものなのか，その両者の定量的な分離が必要である．このためには,「ある学力を有している受験生が，選択していない科目を現在受験している科目に代えてもし選択したとするなら (選択した科目と同等の努力を払ってもし受験するとしたら)，どの位の得点を得るか」についての妥当的な推定値が必要となる．確かに科目選択で選択していないデータを欠測とすることは通常の意味での欠測とは異なっているが，ここでいう欠測は上記のかぎ括弧付きでの意味での欠測ということがいえる．

さて，ここで欠測データの取扱いについて整理しておく．欠測データの取り扱いには以下の 3 つの方法がある．

1. 欠測データを補完 (impute) する．すなわち，欠測値を埋める．
2. 欠測の確率をモデル化する．欠測のメカニズムは Rubin(1976) 以降，以下の 3 つに分けて考えることが一般的になっているが，欠測確率のモデル化はこの順に難しくなる．
   (a) Missing completely at random, MCAR; 欠測するかどうかはモデリングに用いている変数に依存しない．
   (b) Missing at random, MAR; 欠測するかどうかは欠測値に依存せずに観測値に依存する．
   (c) Not missing at random, NMAR; 欠測値は観測していない他の変量にも依存する．
3. 欠測データを無視する．最もよく用いられている listwise deletion と呼ばれる方法は欠測を含む観測値を分析から外すものである．この方法では変量の数が多いと欠測率がさほど大きくなくとも使える観測値の数は極端に少なくなる．ただ相関係数を計算する場合は，そのペアにだけ欠測のない観測値を全て用いる pairwise deletion が使われる．平均の計算ではその変量で欠測したもののみを除くという場合もある．

1. のうち，欠測値にある値を代入することで補完し，擬似的な完全データを作成し，その完全データから解析を行う方法を単一代入法 (single imputation) と呼ぶ．この方法には一般に二つのアプローチがある（より詳しくは Little and Rubin, 2002, pp.59–60).

(i) 明示的 (explicit) なモデリング

予測分布に多変量正規などの統計的なモデルを用いるもので，それ故，仮定が明示的なもの

(ii) 暗黙的 (implicit) なモデリング

補完の焦点がアルゴリズムにあり，モデルの仮定を置かないわけではないが，明らかには示されていないもの．しかしながら補完の方法にはある程度の合理性がある (reasonable) ことが必要である．

(i) の方法には，平均値代入 (mean imputation)，回帰代入 (regression imputation)，確率的回帰代入 (stochastic regression imputation) がある．

(ii) の方法には，ホットデック (hot deck; 同じデータセットの中から最も似た個体の値を欠測値に代入する方法)，代替個体の利用 (substitution; 標本に含まれない別の個体の値を欠測値とし

て置き換える方法)，コールドデック (cold deck; 観測の外部のソースからある一つの定数をもってきてそれで代入する方法；たとえば過去の時系列データから季節調整をした値を用いるなど) がある．

他にこれらの方法のバリエーションとして，例えばホットデックと回帰代入法については予測した値や最も似た個体の値に，データから得られた統計上の誤差をランダムに加えるといった方法などがある（Schieber, 2005; David, 1986）．これは単一代入法が (たとえ補完が適当であったとしても) 推定値の分散を過小に評価してしまうことを改良したものといえる．

この単一代入法を複数回実施し，得られた複数の推定値の統合を行うのが，多重代入法 (multiple imputation; Rubin, 1987) である．この方法は，マルコフ連鎖モンテカルロ法を用いたベイス推定の一種の近似と見なすことができる (星野, 2009).

多重代入法は確かに標準誤差において良い推定量を与えるが，同時に幾つかの制約を要求する．1 つ目はデータが MAR の仮定を満たすこと，2 つ目は行う補完がある意味で正しいこと，3 つ目は分析に用いているモデルが補完に用いられているモデルとある意味で合致していなければならないことである．これら全ての条件についての厳密な記述は Rubin (1987, 1996) にあるが，多重代入法を利用する場合にはこの制約を意識しなければならない．さらに欠測値の補完を行うためには，完全データ（観測データと欠測データ）の確率モデルを置く必要がある．たとえば多変量正規モデル (R では NORM)，対数正規モデル (CAT), カテゴリカル変数に対する対数線形モデルと連続変数に対する多変量正規モデルを結合させた一般位置モデル (general location model; MIX) などがその例である．

2. の欠測の確率モデルについては，現実には MAR の仮定が破られている危険が少なくない．NMAR のモデルとして Multinomial mixture model (Marlin, AISTATS-2005), Aspect model (Hofman, ECML-2001), URP model (Marlin, NIPS-2003) などが知られる．しかしながら，NMAR では仮定したモデルに対しての妥当な推定値を得ることがしばしば難しいため，MAR の仮定を置くことは多い．

いま $Y_{\rm obs}$ と $Y_{\rm mis}$ をそれぞれ Y の観測及び欠測データとし欠測識別行列を M とすれば，MAR の定義により

$$f(M|Y,\phi) = f(M|Y_{\rm obs},\phi) \quad \text{for all } Y_{\rm mis},\phi \tag{1}$$

ただし ϕ は未知パラメータ，であるから観測されているデータと欠測識別変数についての同時分布は，Y の周辺分布 $f(Y|\theta)$ を用いて，

$$\begin{aligned} f(Y_{\rm obs},M|\theta,\phi) &= \int f(Y|\theta)p(M|Y_{\rm obs},\phi)dY_{\rm mis} \\ &= f(Y_{\rm obs}|\theta)p(M|Y_{\rm obs},\phi) \end{aligned} \tag{2}$$

となる (Rubin, 1976)．つまり分布パラメータ θ の最尤推定を行う際には，MAR を仮定すれば欠測識別に関する部分 $p(M|Y_{\rm obs},\phi)$ を無視することができる．

(2) 式は観測されたデータを完全に利用しているという意味で完全尤度（full likelihood; 完全データの尤度とは違う）と呼ばれるが，この完全尤度に基づく推測の方法 (method of full-information maximum likelihood; FIML) はよく知られるように収束しない場合があり，欠測が多いと最尤法の良さがでない可能性が少なくない．また最尤法である以上当然であるが，観測データに対して多変量正規などの統計的確率モデルを仮定する必要がある．

さて，今世紀に入って，Breiman(2001) によって提案された Random Forest(以下 RF と略す）と呼ばれる集団学習の方法がある．この方法は分類や非線形回帰の方法として知られているが，これを用いれば，特別の統計的確率モデルの仮定を置くことなく多くの割合の欠測を含む大量のデータに対して安定してかつ精度のよい欠測データの補完を行うことができる．RF は同じ集団学習の一つである Bagging(Breiman, 1996) を改良したもので，Bagging が全ての変数を用いるのに対し，RF では変数をランダムサンプリングしたサブセットを用いる．このため高次元の解析に向いており，大量データに対して効率的に動作する．ただし欠測のメカニズムに MAR の仮定を置くことは避けることができない．RF では（欠測を含む）全観測データ Y が与えられたもとでの欠測識別と，観測データ Y_{obs} のみが与えられたもとでの欠測識別とを区別しないため，前述の (1) 式が成り立つ．これは MAR の定義そのものである．つまり RF のアルゴリズムそれ自体が，MAR を仮定したデータ補完を行いつつ，補完したデータセットを基に学習を行いそれを繰り返す，というものになっている．したがって最終的な予測モデルを作成するに必要なデータ間距離行列を用いれば，オリジナルのデータセットの欠測値の予測，すなわちデータ補完ができるわけである．RF の現バージョンは Version 5.1, dated June 15, 2004 であるが，明示的に欠測の補完ができるようになったのは Version 4 からである．ただ現時点では，欠測値を補完するためには応答変数（被説明変数）を必要とするモデルでなくてはならない．

RF は遺伝学や生物情報学などの分野等でその有用性が認められつつあり，統計的学習の有名なテキスト “The Elements of Statistical Learning” (Hastie et al., 2009) にも chapter が設けられている．しかし日本語のテキストや解説記事には金 (2006,2007) に数ページの記述があるのに加え，杉本 (2005) に RF の数理的な背景についての報告があるに過ぎないようである．もちろん機械学習やデータマイニングの論文で，対照方法などに用いられている文献はいくつか存在する．しかしながら RF が欠測値の推測に使えること，また多くの欠測値を持つデータの正確さの維持に有効なこと，また補完アルゴリズムの詳細については，執筆時点 (2010 年 11 月 1 日）で日本語に書かれたものはどの記事にも，また Web 上にも存在しないようである．このため，わが国の統計学者や社会統計学者にはおそらくほとんど知られていないが，今後，欠測を多く含む大量データ解析の標準の一つになることは間違いないと確信している．欠測値を補完する方法は，たとえ MAR の仮定を置くにせよ，データの情報を有効に使えるからである．

本稿では欠測の多い大量データ解析の事例として，理科において得点調整が起きそうになったある年度の大学入試センター試験データを用いて，同じ総合学力に対する科目間得点の差を検討する．本事例は機械学習の分野でしばしば用いられる kNN(k-nearest neighbor) 法による欠測値補完では，欠測の割合が大きすぎるために妥当な結果を得ることができないことを予め断っておく．

2 節では RF とその基礎となる CART や Bagging について整理しておく．3 節では RF によるデータ補完の方法について，理論的な側面から説明する．4 節にはある年度のセンター試験の「地理・歴史」「公民」「理科」における科目間得点の差についての解析結果を示す．5 節には全体のまとめと考察を行う．

2. RF を理解するための基礎知識

2.1. CART

RF の基本要素となる CART について簡単に説明しておく．CART(Classification And Regression Tress) は，カリフォルニア大学の Breiman らが 1980 年代初めに公開した樹木モデル (tree-based model) の 1 つである (Breiman, 1984,1998)．非線形分類や非線形回帰に用いられる．CART では説明変数を 2 進分岐させ，2 進木を生成する．初期の CART では分岐の評価基準として経済学者 Gini により提案されたジニ係数 (Gini diversity index, GI) を用いたが，幾つかのバリエーションではエントロピー (entropy) も用いられる．エントロピーとジニ係数の定義を以下に示す．

$$\text{entropy} = -\sum_{i=1}^{c} p(i|t)\log_2 p(i|t)$$

$$\text{GI} = 1 - \sum_{i=1}^{c}[p(i|t)]^2$$

式中の t はノード，i はクラス，p は分割された個体のクラスに属する比率である．エントロピーあるいはジニ係数の最も大きなノードを第 1 ノードとし，以下順に樹木を生成，分岐させてゆく．最近では分岐の評価基準として，C4.5/C5.0/See5 などで用いられる情報利得 (information gain) も好んで使われる．

回帰木の場合は Breiman(1984,1998) では予測値との残差の平均 2 乗誤差，すなわちサイズ N の標本を $(x_i, y_i)(i = 1, \ldots, N)$ とし予測関数を $d(x_i)$ とするとき

$$\frac{1}{N}\sum_{i}(y_i - d(x_i))^2$$

を最小とする．

2.2. Bagging と RF

Bagging は bootstrap aggregating に由来する造語で，Breiman(1996) によって提案された集団学習の方法の 1 つである．そこからも明らかなように Bagging には bootstrap 法が用いられており，その大まかな処理過程は以下の通りである．

いま回帰モデルを考える．訓練データ Z にモデルを適合させる:

$$Z = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

入力データ x における予測関数を $\widehat{f}(x)$ とする．それぞれのブートストラップ標本を $Z^{*b}, b = 1, 2, \ldots, B$ としそこからの予測モデルを $f^{*b}(x)$ とする．Bagging 推定量は

$$\widehat{f_{\text{bag}}}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x)$$

となる．

一方 RF は Breiman(2001) らによって自らが提案した Bagging を改良した新たなデータ解析の方法である．精度や計算資源の点で優れているとされる．Bagging との大きな違いは，Bagging が全ての変量を用いるのに対し，RF では変量をランダムサンプリングしたサブセットを用いるので，高次元データの解析に向いていることである．ランダムサンプリングする変数の数 m は既定値で，Hastie et al.(2009) による推奨値は分類の場合 $\lfloor\sqrt{p}\rfloor$ で回帰の場合 $\lfloor p/3\rfloor$，最小のノードサイズ $n_{\min}$ の規定値はそれぞれ 1 と 5 である．$\lfloor\ \rfloor$ はそれを越えない最大の整数を示す．

RF のアルゴリズムは以下の通りである (Hastie et al. 2009)．なお RF では観測データ (observation) のそれぞれをケース (case) と呼び，変数のことをクラスと呼ぶ．

1. For $b = 1$ to B:
   (a) ブートストラップ標本 Z^* を訓練データから取り出す．
   (b) Z^* からランダムフォレストの木 T_b を成長させる．そのために木のそれぞれの終端ノードに対して，木のノードに割り振られる訓練データの数が $n_{\min}$ に達するまで以下のステップを再帰的に繰り返す．
      i. p 変量からランダムに m 変量を選ぶ．
      ii. m 変数の中から最良の変量と分割点を選ぶ．
      iii. そのノードを 2 つの子ノードに分割する．
2. できた複数の木のアンサンブル (ensemble) $\{T_b\}_1^B$ を出力する．
3. 新しいデータ x に対する予測値を求める．回帰の場合: $\widehat{f_{\mathrm{rf}}^B}(x) = (1/B)\sum_{b=1}^B T_b(x)$ ．
   分類の場合: $\widehat{C_b(x)}$ を b 番目の RF のクラス予測としその多数決をとる（一番多く予測されたクラスの値を選ぶ）．すなわち $\widehat{C_{\mathrm{rf}}^B} = \arg\max |\{(\widehat{C_b}(x)\}_1^B|$ とする．

一般に CART などの樹木モデルでは木の分岐，成長に加えて成長しすぎた木をなんらかの基準に基づいて枝刈りをし，当てはまりのよい簡潔なモデルを構成するが，RF ではこの枝刈りを行わない．RF では仮に 1 つの樹木モデルで過適合になったとしても，ブートストラップ操作による繰り返しを行い，その多数決を取ることで「大数の法則」により最適なモデルが選択できる．

またもし独立同一分布 (i.i.d.) に従う B 個の確率変数の分散を σ^2 とすれば，その平均の分散は $(1/B)\sigma^2$ となる．もし確率変数が同一分布 (独立同一分布でない）に従うとすれば，平均の分散は

$$\rho\sigma^2 + \frac{1-\rho^2}{B}\sigma^2$$

ただし $0 \le \rho \le 1$ は相関の大きさ，となる．もし B が十分に大きければ第 2 項はゼロに近づき，第 1 項だけが残ることになる．つまり集められた木のペアの相関の大きさ ρ にこの値が制限される．これより RF では木の間の相関を小さくすることにより平均の分散を小さくすることがわかる．RF では p 個の変量の中から $m \ll p$ の変量を「ランダム」に選ぶことによりこの ρ の値を小さくしている．

もちろん m を大きくすると個々の木の予測精度を高めるがその一方で木同士の相関も高めてしまう．最適な m の幅があり，このため m は 調整可能なパラメータになっている．Breiman (2001) は，m 個の変数の選び方として 2 つの方法を提案している．一つはこの値を固定する方法で例えば $\lfloor\log_2 p + 1\rfloor$ とする (Forest-RI). もう 1 つの方法は，p が小さいときに選んだ変数間の相関が大きくなることを避けるために, あらたに p 個の中なら ℓ 個の変数を選び，$[-1, 1]$ の一様乱数を係数とした線形結合によって示される新たな変数の中から m 個を選択する (Forest-RC).

たとえば $\ell = 3, m = 2$ とすれば，$2^\ell = 8$ 個の変数の中から $m = 2$ 個の変数を選ぶことができる．

なお Breiman による RF の実装では，交差確認法及び専用のテストデータセットを用いていない．与えられたデータの 2/3 でモデルを作成し，残りの 1/3 データはテスト用のデータ (out-of-bag; oob) として，分類エラーの不偏推定量を得るために使われる．また変量の重要度を得るためにも使われる．

一方，m 個の変数をランダムに選ぶのではなく，用いる変数間の独立性についての帰無仮説を設定し，この仮説が棄却されないような変数を順次作成し，2 分木を成長させる方法もある．この条件付き推論ツリーモデル (conditional inference tree model: cTree; Hothorn, 2006) を RF に適用した方法は cForest(Hothorn, 2007) と呼ばれる．変数の重要度 (variable importance measure) の不偏推定ができるとしている (Strobl, 2008).

3. RF を用いた欠測データの補完

3.1. データ間の近似度

ケース間におけるそれぞれの類似度をある本質的な指標に基づいて測定し，それを $N \times N$ (N は観測データの数；サンプルサイズ）の近似度行列として表現することを考える．この行列は定義により実数の対称行列で，非負である．2 つのケースが互いに類似していなければ 0 を与え，類似していれば最大 1 を与える．そのために以下の手順を行う．最初に近似度行列の全ての要素を 0 とする．次に (木を作成するための）訓練データにおける全てのケースを木に落としてやる．もしケース k と n が同じ終端ノードに落ちたならその近似度を 1 増やす．全てを終え，木の数によって行列全体を割り近似度を正規化する．

巨大データに対しては $N \times N$ 行列をメモリ上に載せることができない場合がある．一つの工夫として，要求されるメモリを $N \times T$（ここで T は木の数）に縮約する．つまり一つのケースの近似度情報を T 次元で表現し，ケース間のベクトル積で近似度を計算する．

3.2. 尺度化

Random Forest の Version 4 から Version 5 への主たる変更点としてこのデータ間の近似行列の計算に改良が計られた．これにより欠測値のより良い補完ができるとしている．手順は以下の通りである．

ケース n と k の近似度を $\{\mathrm{prox}(n,k)\}$ とおく．定義よりこの行列は対称の正定値行列で上限が 1 である．対角要素は 1 となる．$1 - \{\mathrm{prox}(n,k)\}$ はケースの数と等しいかそれ以下の次元をもつユークリッド空間における 2 乗距離である．

ここで　$\{\mathrm{prox}(-,k)\}$ を $\{\mathrm{prox}(n,k)\}$ の第 1 軸にわたる平均とする．同様に $\{\mathrm{prox}(n,-)\}$ を $\{\mathrm{prox}(n,k)\}$ の第 2 軸にわたる平均とし，$\{\mathrm{prox}(-,-)\}$ を両軸にわたる平均とする．

こうすれば行列

$$\mathrm{cv}(n,k) = .5 * (\mathrm{prox}(n,k) - \mathrm{prox}(n,-) - \mathrm{prox}(-,k) + \mathrm{prox}(-,-))$$

は距離の内積を示す行列となり，対称の正定値行列となる．cv の固有値を $\lambda(j)$ とし，固有値ベクトルを $v_j(n)$ とする．このとき

$$x(n) = (\sqrt{\lambda(1)}v_1(n),\ \sqrt{\lambda(2)}v_2(n), \ldots)$$

とすれば $x(n)$ と $x(k)$ の間の 2 乗距離が $1 - \{\mathrm{prox}(n,k)\}$ に一致する．ここで $\sqrt{\lambda(j)}v_j(n)$ は j 番目の尺度化された軸に相当する値である．

この計量尺度化では，最初のいくつかの尺度軸でベクトル $x(n)$ のよい近似を与えるから，cv 行列の最初の大きい幾つかの固有値とそれに相当する固有ベクトルをとるだけで，距離行列のよい近似が得られる．この大きい方から幾つの固有値を用いるかを指定する変数が，Breiman のコードでいうところのそれぞれのケースを表現するのに用いる近似度の数（上位何個までの近似値を用いるか; “retaining only the `nrnn` largest proximities to each case”）である．計算処理を早くするためにサンプルサイズよりかなり小さな値に設定することが推奨されている．

低次元へ落としこんだ射影距離 (projection distances) をより正確に計算する方法としては，たとえば Roweis and Saul アルゴリズムがある．しかしながら次元縮約のための計算速度を優先してこの計量尺度化が用いられている．

3.3. 訓練データにおける欠測データの補完

RF は 2 段階で欠測データを置き換える．最初の段階では連続量に対してはクラス j における全ての変量のメディアンをとりその値に置き換える．カテゴリカル変量に対してはクラス j における最頻値で置き換える．この置き換えは埋め込み (fills) と呼ばれる．

次の段階では，より計算量の多い，しかし良い結果を与える置き換えを行なう．最初の段階での粗い不正確な埋め込み値を用いて RF の木を成長させ，データ間の近似度行列をつくる．

いまデータ行列を $x(n,m)$（n:ケース，m:変数) で表し，これが連続値をとるべき欠測値であるとする．このとき m 番目の変数の値を，この n 番目のケースと (m 番目の変数の値が）非欠測である他のケースとの近似度によって重み付けた平均で埋める．すなわち

$$\widehat{x}(n,m) = \frac{1}{\#\ \text{of non-missing}\ x(\cdot,m)} \sum_{\substack{i \neq n \\ i \in \text{non-missing}}} \mathrm{prox}(i,n)x(i,m)$$

とする．欠測値がカテゴリカル変量のときには非欠測値の最頻値を埋めるが，頻度は近似度によって重みづけたものである．すなわち

$$\widehat{x}(n,m) = \underset{C_m}{\mathrm{argmax}} \sum_{\substack{i \neq n \\ i \in \text{non-missing}}} \mathrm{prox}(i,n)$$

とする．ここで C_m は m 番目のクラスにおける値とする．ここで新しく埋めた値を用いて，もう一度 forest を作る操作を繰り返す．また新しい埋め込み値を用いて繰り返す．Breiman & Cutler (2010) によれば経験上，4〜6 回の繰り返しで十分であるとしている．

3.4. テストデータにおける欠測データの補完

テスト集合の欠測値の補完には, そのテスト集合にラベル（応答変数）が存在するかしないかによって 2 つの異なった方法がある．もしラベルがあるときは，訓練データ集合から導かれる埋め込みを用いる．ラベルがないときはテスト集合におけるそれぞれのケース (個体）ごとにクラスの数（変数の数）の回数だけ更新する．つまり始めにクラス 1 について欠測値を埋め，次にク

ラス 2 について欠測値を埋める．

このようにしてできたテスト集合に対し木を生成する．この更新作業の繰り返しにおいて，ケース（個体）のクラスを決定するために多数決による方法がとられている．

4. センター試験データを用いた解析

4.1. 得点調整

センター試験の本試験について次の科目間で，原則として 20 点以上の平均点差が生じ，これが「試験問題の難易差に基づくものであると認められる」場合には得点調整が行なわれる．

1. 地理歴史の「世界史 B」「日本史 B」「地理 B」の間
2. 公民の「現代社会」「倫理」「政治・経済」の間
3. 理科の「物理 I」「化学 I」「生物 I」「地学 I」の間

得点調整の仕方には分位点差縮小法 (前川, 2001）が用いられ，これは素点 x_j を調整点 z_j に変換するものである．科目間での（最高平均点 − 最低平均点）が 20 点以上のとき，調整点 z_j は次式で与えられる:

$$z_j = wQ(x_j) + (1 - w)x_j$$

ただし

$$w = 1 - \frac{15}{\text{最高平均点} - \text{最低平均点}}$$

とし，$Q(x_j)$ が等百分位による変換である．このとき最高平均点を持つ科目と最低平均点をもつ科目の得点も同じ重み w で変換される．

平均点差で 20 点以上を 15 点に縮小するわけだから，その修正はかなり保守的かつ限定的ではある．しかしながら，平均点差が「試験問題の難易差に基づくものであると認められる」ためには，同じ学力層における平均の科目得点差を論じる必要がある．たとえば「物理 I」の受験者群が他教科の受験者群に比べ相対的に学力が高いといったいわゆる選択バイアスが存在するのであれば，その分の補正や検討が必要となる．従来，このような要望に対する研究については，大津 (2009) が MAR の仮定の下で非線形因子分析を用いて，1 次元の潜在学力分布に正規分布を仮定した場合の 50 個の等分位区間に対する科目間得点差についての比較研究を行なっている．本稿の試みは別のアプローチによる研究である．

4.2. センター試験データへの適用

本節では得点調整が起きそうになったある年度のセンター試験データを用いて同じ総合学力における各科目間の得点比較を行なう．総合学力の定義としては「国語 (200 点）＋数学 I・数学 A（100 点）＋英語筆記 (200 点）＋英語リスニング (50 点)」とする．各科目の偏差値の総和は主成分分析における第 1 主成分に相当するから，これを総合学力とすることに合理性はある．もちろん大津 (2009) の方法も適切である．ただ本研究では取り扱う対象が各科目間の得点差すなわち素点の差であるから，その応答変数（被説明変数）であるところの総合学力も素点である方がわかりやすいのではないかと考えた．これは数学的あるいは統計的妥当性として素点の方が優れているという主張ではなく，むしろ一般の人に向けた理解の容易性を意図したものである．

大津 (2009) の方法に限らず従来法では，欠測データが少なくなるようたとえば理科系科目の解析では「数学 I・数学 A」と「英語（筆記とリスニング)」を受験した受験生を対象に理科の科目の得点比較を，また文科系科目の解析では「国語」と「英語（筆記とリスニング)」を受験した受験生を対象に地理・歴史や公民の科目の得点比較を行なっている．その上で欠測したデータについては MAR の仮定をおいて解析を進めている．しかし理科系大学希望者の中には「国語」を受けないが公民の科目を受けている者は少なくない．センター試験はアラカルト方式を取っており，任意の科目を好きなだけ選択することが可能であるから，ここでの解析の対象から外れた中に，解析すべき科目を受験している者は当然存在しうる．

そもそも得点調整では全受験者を対象とした科目得点の平均差を論じているから，可能な限り捨てるデータを少なくしたいという要請は存在する．(センター試験約 55 万受験者のうち，国立大学を志望する者は約 30 万人であり，多科目受験者だけを解析の対象とすると誤った結論を導いてしまう.）そのための一つの方法が MAR に基づくデータ補完である．もちろん，MAR の仮定が成立している保証のないところに MAR の仮定をおいて欠測データを補完するわけだから，捨てるデータは少なくなる一方で疑わしいデータも多くなる．したがってデータ補完の方が優れているかどうかは必ずしも明白ではなく，それは問題の性質によるとしかいいようがない．ただ一般論として，欠測率が高くなければ従来捨てていたデータを取り入れた方が全体としての推測の精度はあがるであろう．そのために（欠測率の高いデータ対象に補完を行なわないようにするために)，著者らは 2 段階でデータ補完を行なう．

1. 最初に（「科目」ではなく）「教科 (試験枠)」単位でデータを収集し，欠測値に対してデータ補完を行なう.「教科 (試験枠)」とは「国語」,「地理・歴史」,「公民」,「数学①」,「数学②」,「理科①」,「理科②」,「理科③」,「外国語」,「英語リスニング」をいい，これらを説明変数とし,「総合学力（国語＋数学 I・数学 A ＋英語筆記＋英語リスニング)」を応答変数（被説明変数）とする回帰モデルを RF でモデル化し，欠測値を埋める．(総合学力においてはそれに含まれる科目の 1 つでも欠測であるならば，その総合学力は欠測であるとする.）
2. 次に着目している教科（「地理・歴史」,「公民」,「理科」）において得点調整の対象となる科目ごとの得点に分け欠測値を埋める．たとえば「地理・歴史」において「世界史 B」「日本史 B」「地理 B」の 3 つに分け欠測値を埋める（この 3 科目は高々1 科目しか受験することができないので，少なくとも 2 科目は欠測値である．）このとき「公民」や「理科」については教科を分けることをしない．これによりデータ全体に対する欠測の割合の低下を必要最低限に抑えることができる．RF における回帰モデルでは「世界史 B」「日本史 B」「地理 B」を応答変数（被説明変数）とし，この結果に基づき「世界史 B」「日本史 B」「地理 B」の得点差を検討する.（同じように「公民」「理科」について実行する.）

4.3. 科目間比較

「地理・歴史」における科目間比較を行なうために，総合学力を横軸にとり，「世界史 B(Wrd_history)」「日本史 B(Jpn_history)」「地理 B(Geography)」の各科目の得点を縦軸に示したのが図 1 である．これより総合学力 300 点の地点で見ると「地理 B」のスコアが他の 2 科目に比べ 5 点ほど高い．総合学力が高くなるにつれこの差はなくなり，総合学力 400 点で 3 者はほぼ一致する．総合学力 500 点の高学力層にとっては「世界史 B」が最も高いスコアを示している．

つまり中学力層においては「地理 B」が良いスコアを取りやすく，高学力層においては「世界史 B」が良いスコアを取りやすいことがわかる．

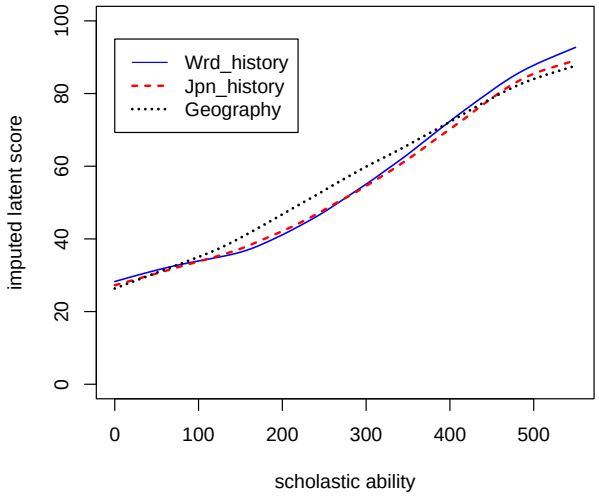

図 1. 「地理・歴史」における科目間比較

同様に「公民」において横軸に総合学力，縦軸に「現在社会 (Cotmpr_soc)」「倫理 (Ethics)」「政治経済 (Poly&Econ)」のスコアをとって示したのが図 2 である．

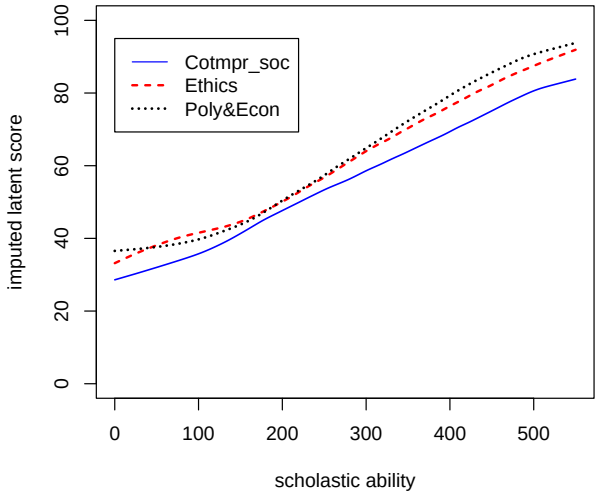

図 2. 「公民」における科目間比較

これより全ての学力層において「現在社会」が他の 2 科目より 5〜8 点ほど少なく，難しかったことが見てとれる．

つぎに「理科」において横軸に総合学力，縦軸に「生物 I(Bio)」「化学 I(Chem)」「物理 I(Phys)」「地学 I(Earth)」のスコアをとって示したのが図 3 である．

総合学力 300 点の中位のあたりで「化学 I」と「地学 I」の間には 20 点近くの違いがあり，その差は総合学力 500 点の高学力層においても変わらないことがわかる．つまり中上位層において「化学 I」が易しく「地学 I」が難しかったことがわかる．

実受験者のデータのみに基づく「化学 I」と「地学 I」の平均の差は約 18 点であり，得点調整を行なう一歩手前まできていたわけだが，この差は受験者の学力の違いによるものではなく，試験の難易差の違いにほぼ負うものであることが確認された．このことは大津 (2009) の解析とほぼ同じ結論である．ただしこの傾向は毎年ごとの科目間の問題の難易によって変わるもので，この年度に限った現象であることを明記しておく．

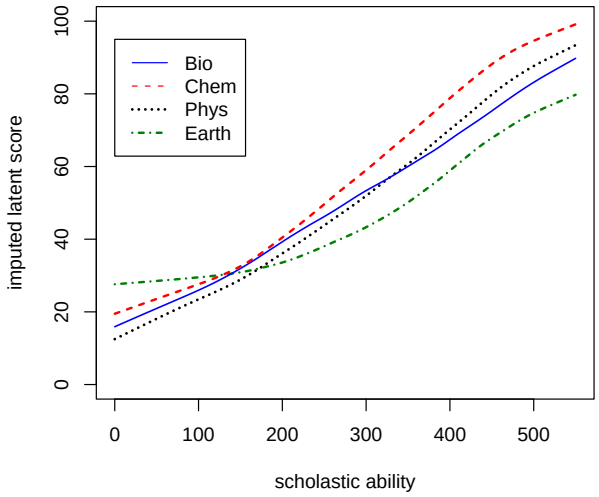

図 3. 「理科」における科目間比較

なお，分析の対象とした科目受験者の数を表 1 に示す．

表 1. 解析の対象とした年度の本試験科目受験者数

地理・歴史	受験者数	公民	受験者数	理科	受験者数
世界史 B	94,106	現代社会	169,711	生物 I	176,043
日本史 B	144,327	倫理	53,116	化学 I	200,411
地理 B	109,616	政治・経済	82,804	物理 I	143,646
				地学 I	25,921

4.4. 補完データと実データとの差異

欠測を補完したデータと実測値との同じ総合学力に対する差について検討することは興味深い事項である．横軸に総合学力，縦軸に欠測の補完データ (Imputed) のスコアと実測データ (Actual) のスコアを理科 4 科目について図 4 に示す．

実測データが補完データよりも（同じ総合学力において）高スコアならば，それは本来その学力から期待されるスコアよりも実際に受験したスコアの方が高いわけだから，その人にとっての得意科目であるといえる．

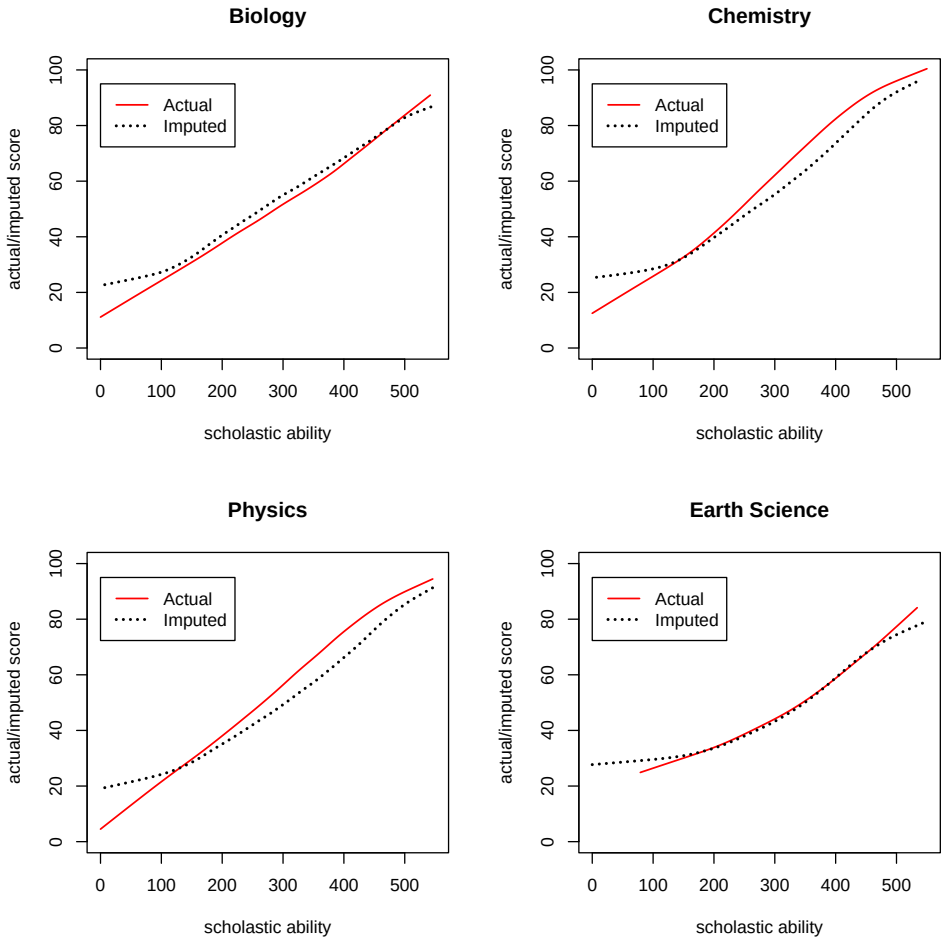

図 4. 「理科」における科目間比較

これより総合学力が 150 点以上の，すなわち大学受験を目指すほぼ全ての受験生において,「物理 I」と「化学 I」の受験者はそれを得意科目にしている（あるいは得意とする人が受験している）ことがわかる.「生物 I」においては，その逆で受験者はそれを不得意科目にしている（あるいは

不得意とする人が受験している）ことがわかる.「地学 I」においては得意・不得意の違いはないように思われる.

「地理・歴史」「公民」においても同様のグラフを作成しているが，全ての科目において得意・不得意の違いは現れなかった.

青木, 他 (2010) にもあるように，志望先における学部・学科と選択科目については明確な関係があり，生物は文系，とりわけ法学系の受験生が受験する科目であり，医学系は物理，化学，生物の 3 科目を，理工系は物理，化学の 2 科目を受験する．理科においては「地理・歴史」「公民」と異なり，文理の志望の別が科目選択に直接的に影響を与える，言い替えれば選択バイアスの問題があり，この点を踏まえた得点調整の必要性がこのデータからも示唆される.

もちろん，このようなことが言えるためには，欠測値の補完が妥当に行われているということが前提となるわけだが，それは図 1-3 の総合学力に対する科目間得点差の違いの傾向や程度が，実測データよりも補完データが多いにもかかわらず，大津 (2009) の解析とほぼ同じであることからも示されると考える．加えて，もしここでの欠測の補完が MCAR(Missing completely at random) に基づくものであったなら，補完データと実データとの差異は生じない．選択バイアスの問題が特徴的な理科においてのみ，一般に納得しやすい結果（「物理 I」と「化学 I」の受験者はそれを得意科目にしており「生物 I」の受験者はそれを不得意科目にしている）が導かれたことは，その妥当性の証左であろう.

5. おわりに

本稿は論文としての格段の新規性はなく，あまり知られていないとはいえ，ただ単に RF を欠測を含む大規模な実用データに適用しそのデータ補完の妥当性を示しただけである，という謗りは甘受せざるを得ないであろう．しかしながら，本事例では教科の中に科目が属するという関係があることを利用し，各科目全体を同列の説明変数として取り扱うのではなく，いったん各教科を説明変量として総合学力を応答変数（被説明変数）とするデータ補完を行ない，次に着目している教科（「地理・歴史」,「公民」,「理科」）において得点調整の対象となる科目ごとの得点（たとえば「地理歴史」における「世界史 B」,「日本史 B」,「地理 B」の得点）における欠測値を応答変数（被説明変数）とするモデルで埋めるという 2 段階の補完という工夫を行っている.「世界史 B」を選択した者は「日本史 B」や「地理 B」を同時に受験することはできず，したがって全科目を同時に受験する者は一人としていない．このようなデータに対して全科目を説明変数として RF によるデータ補完をすることはそもそも実行不能なのであるが，本稿のような工夫をすれば実行可能となる．この事例のように，説明変数が全て同等もしくは同列ではなく，幾つかの説明変数がグループにまとめられ，またそのグループの中から一つが排他的に選択されるような場合には本稿の手順は有効であろう.

また RF は高次元でかつ欠測が少なくないデータに対して比較的頑健に実行でき，これによりデータ構造の理解に有効に機能することはよく知られた事実であるが，本稿の事例においても同様であることが改めて確認された.

RF によるデータ補完は本稿のような 2 段階の補完をしない場合であっても，1 節で述べた欠測データの取り扱い方法の分類の中では，単一の推定を何度か行い得られた複数の推定値を統合するという点で，多重代入法と似たアプローチであるといえる．両者の違いは RF によるデータ

補完がデータモデルに恣意的な確率モデルをおく必要がないことと，データ間の近さを分類木における最終ノードの位置を問題としているという点であり，これは統計的なアプローチとはかなり様相が異なっている．両者の相違性や近似性についての数学的な検証が今後行なわれることを期待している．

また現在，自然言語処理分野を中心とする機械学習の分野では半教師学習 (semi-supervised learning) と呼ばれる，ラベル付けされた（応答変数の値がわかっている）データ集合に加え，ラベルのない（応答変数の値がわからない）データ集合を含む「ラベルあり・なし混在データ (labeled and unlabeled data)」 から学習することで，ラベルありデータだけで学習した場合より，より予測精度の高いクラス分類を実現する方法についての研究が盛んになっている．この方法との関連についても今後の研究が期待される．

謝　辞

本論文について貴重なコメントをいただき，また多くの不備や幾つかの誤りについて御指摘くださった 2 名の匿名の審査員に心より感謝します．また同様に本論文について有益なコメントをいただきました繁桝算男先生（帝京大学）, 岸野洋久先生（東京大学），大津起夫先生（大学入試センター），櫻井裕仁先生（大学入試センター）にお礼申し上げます．

なお本研究の一部は基盤研究 (B)「試験問題統計情報のデータベース化と自然言語処理技術を用いた統計解析」の一環として行われ，科学研究費補助金により助成されています．ご支援をここに深く感謝します．

補 遺 ： R プログラムによる欠測値の補完

RF の詳しい使用法については金 (2007) 他を参考にされたい．判別・回帰における変数の重要度としてのジニ係数の表示の仕方についての言及もここにある．RF による欠測データの補完については，CRAN より入手できる同じ `randomForest` ライブラリにある `rfImpute()` を使う．

R ライブラリには他に `yaImpute()` なる kNN(k-nearest neighbor) によるデータ補完アルゴリズムがある (Crookston, 2008)．著者らはこのライブラリによるデータ補完 ($k = 1$) を用いたが妥当な推定値を得ることができなかった．データの次元の数が大きくなると，いわゆる「次元の呪い (curse of dimensionality)」によってどのデータも互いに似なくなる．このため多次元のままで，かつ欠測率が大きいデータに単に kNN を適用しただけでは，データのもつ本質を捉えることが難しく，少なくとも本事例ではうまくいかなかった．

読者の利便のために，有名な Fisher のアヤメのデータを用いた `rfImpute()` によるデータ補完のプログラム例をここに示す．データに恣意的に欠測（4 つの説明変数に対しそれぞれ 20/150 の欠測）を与え，その欠測を `rfImpute()` で補完し，補完データに対して，再度 `randomForest()` で応答変数 (`Species`) を推測している．

```
> library(randomForest)
> data(iris)
> iris.na <- iris
> set.seed(111)
> ## artificially drop some data values.
> for (i in 1:4) iris.na[sample(150, sample(20)), i] <- NA
> set.seed(222)
```

```
> iris.imputed <- rfImpute(Species ~ ., iris.na)
> set.seed(333)
> iris.rf <- randomForest(Species ~ ., iris.imputed)
> print(iris.rf)

Call:
 randomForest(formula = Species ~ ., data = iris.imputed)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 5.33%
Confusion matrix:
           setosa versicolor virginica class.error
setosa         50          0         0        0.00
versicolor      0         46         4        0.08
virginica       0          4        46        0.08
>
```

参 考 文 献

青木敏，大津起夫，竹村彰通，沼田泰英 (2010): 大学入試センター試験科目選択データの統計解析，応用統計学 **39**，[2&3]，71–100.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone., C. J. (1984): Classification and Regression Trees, Wadsworth.

Breiman, L. (1996): Bagging Predictors, Machine Learning, **24**, 123-140.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1998): *Classification and Regression Trees*, Chapman & Hall/CRC.

Breiman, L. (2001): Random Forests, Machine Learning, **45**, [1], 5-32. (DOI: 10.1023/A:1010933404324, `http://www.springerlink.com/content/u0p06167n6173512/`)

Breiman, L. and Cutler, A. (2010): Random Forests, `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`

Crookston, N.L. & Finley, A.O. (2008): yaImpute: An R Package for *k*NN Imputation, Journal of Statistical Software **23**, Issue 10. `http://www.jstatsoft.org/v23/i10`

David, M., Little, R.J.A., Samuhel, M.E. & Triest, R.K. (1986): Alternative methods for CPS income imputation. J. Amer. Statist. Assoc., **81**, 29-9641.

Hastie, T., Tibshirani,R., & Friedman, J. (2009): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

星野崇宏 (2009): 調査観察データの統計科学-因果推論・選択バイアス・データ融合，岩波書店.

Hothorn, T., Hornik, K., & Zeileis, A. (2006): Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**[3], 651-674.

Hothorn,T., Zeileis, A., & Hornik, K. (2007): Let's Have a party! An Open-Source Toolbox for Recursive Partytioning. Research Report Series **59**, Deptartment of Statistics and Mathematics, Wirtschaftsuniversität Wien.

金明哲 (2006): R と集団学習, ESTRELA 2006 年 3 月 (No.144), 64–70.

金明哲 (2007): R によるデータサイエンス, 森北出版.

Little, R.J.A. and Rubin, D.B. (2002): Statistical Analysis with Missing Data, 2nd edition, New York: John Wiley.

前川眞一 (2001): 大学入試センター試験における選択科目間の得点調整について, 計測と制御, **40** [8], 568-571.

大津起夫 (2009): 平成 20 および 21 年度大学入試センター試験における科目別得点の難度比較, 大学入試センター研究開発部リサーチノート, RN-09-11(内部資料).

Rubin, D.B. (1976): Inference and missing data. Biometrika, 63, 581–592.

Rubin, D.B. (1987): Multiple imputation for nonresponse in surveys. New York: Wiley.
Rubin, D.B. (1996): Multiple imputation after 18+ years (with discussion). Journal of the American Statistical Association, 91, 473–489.
Schieber, S.J. (2005): "A comparison of three alternative techniques for allocating unreported social security income on the survey the low income aged and disabled." American Statistical Association, Proceedings of the Section on Survey Research Methods, 212–218.
Strobl, C. & Zeileis, A. (2008): Exploring the statistical properties of a test for random forest variable importance. In COMPSTAT 2008 - Proceedings in Computational Statistics. Volume II. Physica Verlag, Heidelberg; 59-66.
杉本知之, 下川敏雄, 後藤昌司 (2005): 樹木構造接近法と最近の発展, 計算機統計学 **18**(2), 123-164.
Yan He (2006): Missing data Imputation for Tree-Based Models, A dissertation for the degree Doctor of Philosophy in Statistics, University of California.

（2011 年 1 月 22 日受付　7 月 8 日最終修正　8 月 5 日採択）

著者連絡先：〒 153–8501 東京都目黒区駒場 2–19–23
大学入試センター 研究開発部
石岡 恒憲
E-mail: tunenori@rd.dnc.ac.jp

Japanese J. Appl. Statist. 40 (3) (2011), 193–210

Data imputation by Random Forest — The principle and its application for National Center Test in Japan —

Tsunenori Ishioka

The National Center for University Entrance Examinations

Abstract

Random Forest, one of the ensemble learning methods for classification and non-linear regression model, provides a stable and an accurate data imputation for the missing data. This paper shows that the algorithm works well for a large dataset containing missing data. The examples are science and society examination scores appearing in the Japanese National Center Test in 200x.

Key words: imputation, missing data, missing at random, ensemble learning

E-mail address: tunenori@rd.dnc.ac.jp (Tsunenori Ishioka)

Received January 22, 2011; Received in final form July 8, 2011; Accepted August 5, 2011.