Random Forestsを用いた欠測値補完の半教師データおよび教師なしデータへの拡張

石岡 恒憲
大学入試センター 研究開発部

tunenori@rd.dnc.ac.jp

1　はじめに

Random Forests [2]（以下RFと略す）は，L. Breimanらによって提案されたCARTなどの分類木を集団学習する比較的新しい方法で，分類や回帰の方法として広く知られている．また，半教師データとは，応答変数が完備しているいわゆる「教師ありデータ」と，応答変数の不明な「教師なしデータ」が混在するもので，このデータを用いることで，「教師ありデータ」のみから学習する「教師あり学習 (supervised learning)」に比べ，その予測精度を向上することができる．「教師なしデータ」は「教師ありデータ」に比べ一般に安価にかつ大量に入手することができ，そのために入手したデータで現実世界をかなりの程度，被覆することが期待できる．従来の統計手法では現実世界を再現できるよう，その標本抽出に細心の注意が払われてきたわけだが，大量の「教師なしデータ」を利用することで，抽出についてのバイアスや，得られた標本についての恣意的なモデルの仮定を避けることができる．

2　RFを用いた欠測値補完

「半教師データ」それ自体が応答変数の欠測データであるわけだが，ここでは予測変数の欠測の補完について議論する．「教師なしデータ」についても同様とする．RFそれ自体は，この半教師データを学習データとする「半教師あり学習 (semi-supervised learning)」を行うことはしていない．RFのBreimanによるFortran77の最新コードは2004年のVer5.1であるが，Ver.4から「教師ありデータ」における「分類問題」に対して，予測変数に欠測値が含まれていてもそれを適当に補完し動作するようになった．補完の方法には2つのオプションがある．一つはmissquick(Ver.4)で，連続値においては，欠測値をそのカラムの非欠測値のメディアンで埋め，離散値においては最頻値で埋めるものである．もう一つの方法はmissright(Ver.5)で，missquickから始めて，ケース間の近似度を用いて欠測値をより正確に置き換える操作を繰り返すものである．欠測値は，非欠測値の近似度に応じた重み付け総和によって埋められる．この考えにしたがって，Andy Liaw[6]は，統計言語Rによって，欠測値を補完する関数rfImputeを実装した．しかし，Breimanのアイディアは，分類から回帰への適用だけでなく，(応答変数の存在しないいわゆる)「教師なしデータ」や（応答変数の一部，通常の場合，その多くが欠測する）「半教師データ」に対しても，ケース間の近似度が計算できさえすれば，適用することが本来は可能である．本稿では，そのための拡張について紹介する．

3　提案する新しい手順

3.1　前提

欠測値の補間に必要なのは，ケース間の距離である．RFではこの距離は，分類木において同じノードに落ちる確率で求めている．RFにおける分類木作成の標準的な（既定の）繰り返し数は500であるので，その試行の結果から当該確率を求める（RFでは一部の変数のみを用いて木を作成するので、この確率は決して小さくはない）．これよりRFに基礎を置く欠測値補完には，応答変数が必須で，またその応答変数は分類データでなくてはならない（回帰では適用できない）．

3.2　半教師データの場合

提案する手順はきわめてシンプルである．最初に予測変量 x に対し粗い補完を行い，RFを用いて欠測の応答変量 $\hat{y}$ を補完する．次にケース間近似度を用いて $\hat{x}$ を推定し，改めて $\hat{y}$ を求める．これ以上の繰り返しは行わない（繰り返しによって推定値の改善をもたらさない）．この手順のキーポイントは，欠測値の $\hat{x}$ にはケース間の距離（近似度）を用いるが，欠測値の $\hat{y}$ に

は，ケース間の距離ではなく，RF モデルに基づく推定値を用いるところにある．

このアルゴリズムの R による実装は，著者の Web ページ [4] の `rfImputeSmsupvsd.R` にある．プログラム中，入力値は予測変数を示すデータ・フレームあるいはデータ行列である x と，応答変数を示すデータ・フレームあるいはデータベクターである y である．欠測値補完を考えているので，x には，当然，欠測 (NA) を含み，y には欠測 (NA) を含んでいてもよい．また繰り返しを示す変数 `iter` を引数として与えるが，標準では繰り返しをおこなわないので，既定値を 1 にしてある．

3.3 教師なしデータの場合

教師なしデータでは応答変数が存在しない．そのため RF では，存在しない応答変数を作成するというトリックを用いる．最初に与えられたデータに対して，応答変数の値をすべて 1 とする．次に応答変数の値が 2 のデータを人工的に作成する．人工データの作り方には 2 通りの方法があり，プログラム中のパラメータで指定することができる．

1. 各変数の周辺分布が一致するように（変数ごとに独立にブートストラップ法を用いて），各変数データをサンプリングする．
2. 各変数の最大値と最小値を範囲とする多次元立方体 (hypercube) の中から乱数を用いてデータを作成する．

この人工データの作成により教師なしデータを 2 値の応答変数を持つ分類問題として書きなおすことができ，これにより各ケース間の距離の算出ができる．欠測値の埋め込みは 2 段階でおこなう．最初の段階で予測変量 x に対し粗い補完を行い，次の 2 段階目で，構成した RF モデルから得られるケース間近似度を用いて $\hat{x}$ を推定する．ここで我々の方法のキーポイントは，距離によるデータの重みづけに学習データ（ケース）の全てを用いるのではなく，最寄りの k 個のケースのみを用いることである．この方法は Liaw の rfImpute と k 最近傍法の折衷といえる．一種の刈り込み平均を用いるのと同じ理屈により，異常値に対してロバストで好ましい結果を与えることが我々の数値実験により確かめられている．カテゴリカルな予測変量に対しては k 最近傍の与える値の多数決により決定する．この 2 段階目の操作を通常は 4–6 回繰り返す．

このアルゴリズムの R による実装は，著者の Web ページ [4] の `rfImputeUnsupvsd.R` にコメント入りで公開されている．メイン関数である `rfunsupvsd` の引数である x は教師なしデータ・フレームあるいはデータ行列である．当然ながら欠測 (NA) を含む．応答変数は不要である．`iter` は反復回数であり，既定値として 5 が与えられている．

4 数値例

本実験では，機械学習でよく用いられるヒューレット・パッカード研究所のスパム/ノンスパムの判別データ [7] と，統計学の分野で非常によく知られたフィッシャーのアヤメのデータ [1] を用いる．スパム/ノンスパムの判別データは 4601 件のデータで 58 変量からなる．最初の 48 変量は e-mail に含まれる文字列（たとえば “business”) の頻度を示す．49–54 番目の変量は “;”, “(”, “[”, “!”,“$”, “#”の各記号の頻度である．55–57 番目の変量は大文字で書かれた文字列長さの平均値，最大値，総計である．58 番目の変量は “spam” か “nonspam”か，すなわち，求められていない商用のメールであるか否かを示す．“spam”が 1813 件で，“nonspam”が 2788 件である．

フィッシャーのアヤメのデータは，3 種類のアヤメ種類（“Iris setosa,” “versicolor,” and “virginica”) のそれぞれ 50 個について,「がく (sepal)」の長さと幅,「花弁 (petal)」の長さと幅をセンチメーター単位で測定したデータである．統計言語 R では “iris”が，150 件の 5 変量からなるデータ・フレームとして予め用意されている．このデータを多次元尺度法 (multidimensional scaling: MDS) で 2 次元布置を行えば，“Iris setosa”は比較的よく分離できるが，“versicolor”と “virginica” については，その分離が難しいことが見てわかる．

4.1 半教師データの場合

応答変数 y，すなわちスパム/ノンスパムの判別データでは 58 番目の変量，アヤメ・データでは 5 番目の変量も他の説明変量 x と同じようにランダムに欠測させる．欠測率は 5%, 10%, 20%, 30%, 40%, 50%, 60%とする．評価すべきは欠測させた y が正しく補完＝推定された程度，すなわち正判別率である．

提案する手法 “rfImput.smspvsd”の性能を例証するために，本手法を 2 つの従来法と比較する．

1. Liaw’s “rfImpute” [6]:この関数は本来，応答変数 y の欠測を想定していない．これは “randomForest”それ自体がそうであるからである．そこで，非欠測の y を用いてそれに対応する x を補完し，フォレストモデルを作成．欠測の y に対しては，

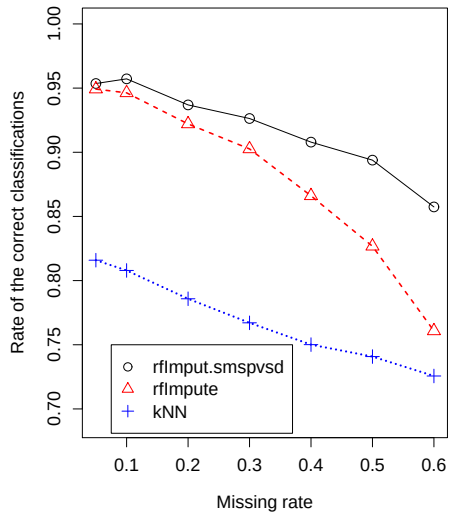

図 1: 半教師あり spam/non-spam データにおける正判別率

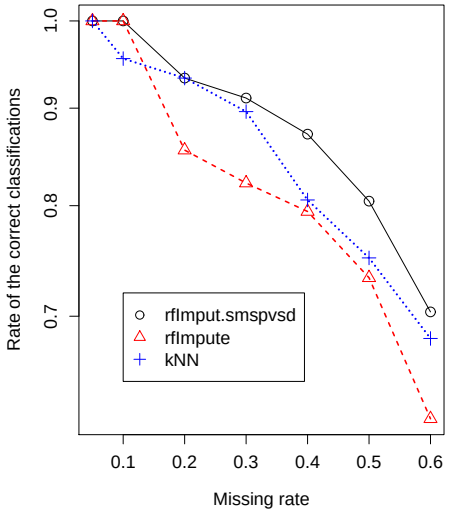

図 2: 半教師ありアヤメ・データにおける正判別率

このモデルとそれに対応する粗い補完を行った $\hat{x}$ から $\hat{y}$ を推定する．

2. “kNN” [5]: 非欠測の y に対する粗い補完を行った $\hat{x}$ から訓練モデルを生成する．欠測の y に対しては，このモデルとそれに対応する粗い補完を行った $\hat{x}$ から $\hat{y}$ を推定する．$k = 3$ とする．

スパム/ノンスパム・データに対して，比較の 3 方法について，それぞれ 3 回の試行を行い，正分類率の平均値をプロットしたのが図 1 である．横軸が欠測率で，縦軸が正判別率である．縦軸の値が大きいほど，良い性能であることを示す．縦軸の値 1 は，y の全ての欠測値が正しく補完されたことを示す．

一般に欠測率が大きくなるにつれて，正判別率は低下する．ランダムにデータを欠測させたために，折れ線は必ずしも単調減少していないが，我々の方法 (“rfImput.smspvsd”) は 3 方法の中で，欠測率によらず最良であることがわかる．kNN は 3 つの方法の中で，明らかに悪い．これは高次元においては，いわゆる「次元の呪い (curse of dimensionality)」により，互いのデータが似なくなることによる．我々の方法は，欠測率が高い（たとえば 60%) 場合において，その良さは顕著である．

同様にアヤメ・データに対して行った結果が図 2 である．このデータも，欠測のランダム化の影響によって，折れ線が必ずしも単調減少していないが，我々の方法が 3 方法の中で，欠測率によらず最良であることがわかる．

一般に，半教師データでは，わずかのラベル付きデータに加え，より多数のラベルなしデータを用いて，推測をおこなう．これは，この例では y の欠測率が大きいことを意味する．このため，y の欠測率を 90%とし，x の欠測率を 5〜 60%とした同様の実験を行っている．結果については紙面の都合で割愛せざるを得ないが，我々の方法が 3 方法の中で，欠測率によらず最良であるだけでなく，x の欠測率が低下しても，正判別率の低下の程度が少ないことが確認されている．たとえば x の欠測率が 60%(y の欠測率が 90%) において，正判別率は 83%程度を確保する．

4.2 教師なしデータの場合

提案する手法 “rfImput.unspvsd” の性能を例証するために，本手法を 2 つの従来法，すなわち “na.roughfix” と “impute.knn” と比較する．前者は我々の方法のベースラインとするもので，欠測値を各変量における中央値で補完するものである．後者は R の biocLite ライブラリーに格納されている kNN 法 [3] によるものである．これは教師なしデータについて欠測値補完を行うものである．近傍 k は，本関数の既定値である 10 とする．

図 3 は，スパム/ノンスパムの判別データによる結果である．応答変数である 58 番目の変量は使わずに，それ以外の 57 変量について，ランダムにデータを欠測させた．欠測率は 5%, 10%, 20%, 30%, 40%, 50%, 60%とし，横軸に目盛る．縦軸は補完した値と本来の正しい値との（相対）残差 2 乗和である．したがって

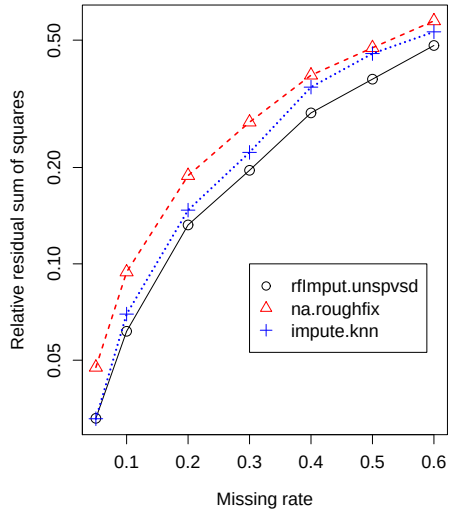

図 3: 教師なし spam/non-spam データにおける残差 2 乗和

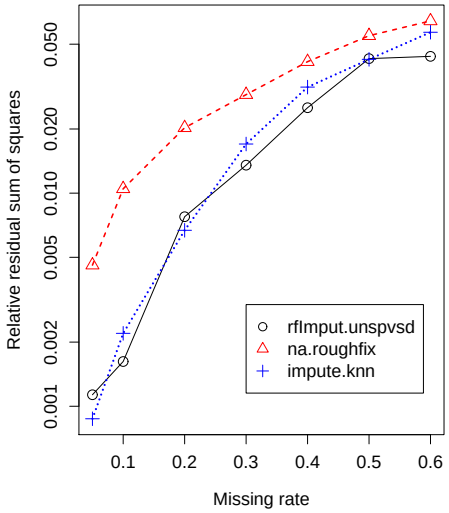

図 4: 教師なしアヤメ・データにおける残差 2 乗和

値が低いほど良い推定＝補完が行われていることを示す．欠測率が上がるにつれ，推定の誤差は大きくなるのは当然である．

3 方法を比較すると，我々の方法が欠測率によらず，最も優れていることがわかる．粗くいって，我々の方法は，ベースライン法である “na.roughfix” を 20-30%，kNN 法を 5-10%程度，向上させる．

もう一つの例は，アヤメ・データに対してである．スパム/ノンスパムの例と同じ枠組で，横軸に欠測率を，縦軸に（相対）残差 2 乗和を示したのが図 4 である．応答変数である 5 番目の変量は使用しない．折れ線が多少がたがたするのは，欠測をランダムにおこなったことと，4 変量 150 個というサンプルサイズの少なさによるものと思われる．我々の方法は，ベースライン法である “na.roughfix” を欠測率によらず，かなりの程度，改善することがわかる．

5 まとめ

スパム/ノンスパム判別データとアヤメ・データを用いた結果，k 最近傍法など既存の従来法に比べて，教師なしデータにおいては推定値の残差二乗誤差を，半教師データにおいては，正判別率（正しく分類推定できる割合）を向上させることがわかった．この方法は，欠測のデータをいったん粗い補完を行い，それを用いて補完値を逐次改善していくという点で，統計学でいうところの多重代入法の一種とみなすことができる．決定的な違いは，多重代入法がデータ間の近さに，(統計的な処理をしたうえでの）空間的な距離を使うのに対し，本手法は木構造モデルにおいて同一ノードに落ちる確率で定義しているところにある．また統計的な手法では，データ構造にあるモデルを仮定するという点で，推定の精度に影響を与える恣意的なパラメータが多いのに対し，本手法で設定すべきパラメータは，せいぜい，Random Forests における木の数や，反復の数といったあまり影響度の大きくないもので，恣意性の少なさがその特徴といえよう．

参考文献

[1] Becker, R. A., Chambers, J. M. and Wilks, A. R.; The New S Language. *Wadsworth & Brooks/Cole*, 1988.

[2] Breiman, L. and Cutler, A.; Random Forests, `http://www.stat.berkeley.edu/~breiman/RandomForests/`, updated Mar. 2004.

[3] Hastie, T., Tibshirani, R, Narasimhan, B. and Chu, G.; impute: impute: Imputation for microarray data, R package version 1.14.0

[4] Ishioka, T.; R functions to impute Unsupervised /Semi-supervised Data using the proximity from Random Forests, `http://coca.rd.dnc.ac.jp/tunenori/rfImpute.html`

[5] k-Nearest Neighbour Classification, R Documentation, knn {class}, `http://stat.ethz.ch/R-manual/R-patched/library/class/html/knn.html`

[6] Liaw, A. and Wiener, M.; *Classification and Regression by RandomForest*, R News Vol. 2/3, 18-22, 2002.

[7] UCI Repository of Machine Learning Databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`