5ZC-06

Random Forest を用いたコマンド履歴からのプログラミングスキル推定

清野真理子 †　　橋本玄基 ‡　　大枝真一 ‡
木更津工業高等専門学校 制御・情報工学専攻 †　　木更津工業高等専門学校 情報工学科 ‡

1. まえがき

近年，スマートフォンや Web アプリケーションなどの IT 技術の発展とともに IT 需要が高まり，国内では IT 技術者不足が深刻になっている．そこで，教育機関では IT 技術者の育成に力を入れている．

効果的な教育を行うには，プログラマのスキルを定量的に評価する必要がある．しかし，プログラミングのスキルは，完成したプログラムを人が読んで採点せざるを得ない．それには時間などのコストがかかる．また，プログラマがプログラムを完成させるまでにかかった時間，修正した回数などは考慮されていない．

そこで，本研究ではプログラミング中のコマンド使用履歴に着目し，Random Forest を用いてログデータの解析を行う．これにより，プログラミングスキル推定を行うことを試みる．

2. 決定木

パターン認識において，木構造は重要なモデルである．木構造の中でも，分類や回帰に用いられる木は決定木 (decision tree) と呼ばれ，特に分類問題で用いられる木を分類木，回帰問題で用いられる決定木を回帰木と呼ぶ．決定木はクラスや目的変数の予測だけでなく，重要な属性やその値の範囲などを解析するためにも使用することができる [1].

決定木には，2 分木と多分木がある．2 分木では CART(classification and regression trees)[2] が良く知られており，多分木では C4.5[3] が良く知られている．

2.1 分類木と回帰木

特徴ベクトル $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T, (x_i \in \mathcal{R})$ および目的変数 y について考えるとする．y がクラスを表せば分類問題，実数であれば回帰問題となる．図 1 に，2 次元の特徴ベクトル $(d = 2)$ を持つデータの例を示す．

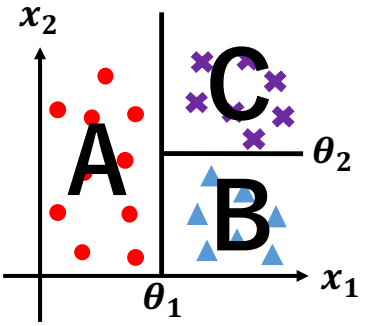

図 1　2 次元の特徴ベクトルを持つデータの例.

図 1 に示すデータは，クラス $c_1 = A$, $c_2 = B$, $c_3 = C$ の 3 つのクラスが存在しており，横軸は特徴量 x_1，縦軸は特徴量 x_2 を示している．丸はクラス A，三角はクラス B，バツはクラス C のデータである．

このデータは，閾値 θ_1 および θ_2 を用いて分類することができる．$x_1 < \theta_1$ の時クラス A，$x_1 \geq \theta_1$ かつ $\theta_2 < x_2$ ならクラス B，$x_1 \geq \theta_1$ かつ $\theta_2 \geq x_2$ ならクラス C となる．

このように，特徴量を閾値で分けることによってデータのクラスが決定される場合，分類木を用いる方法が有効となる．図 1 を分類木で表したものを図 2 に示す.

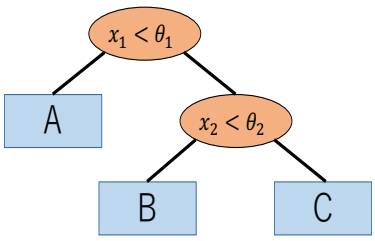

図 2　図 1 の分類木.

図 2 において，丸や四角のことをノードといい，丸を中間ノード，四角を葉ノードという．中間ノードは条件を，葉ノードはデータのクラスを表す．最も上にあるノードを根ノードといい，また，あるノードの下にあるノードを子ノードという．

決定木は，中間ノードで閾値未満なら左の子ノードに，閾値以上なら右の子ノードにデータを分類していく．

Estimation of Programming Skill from Command History using Random Forest
†Mariko Seino・National Institute of Technology, Kisarazu Collage
‡Genki Hashimoto, Shinichi Oeda・National Institute of Technology, Kisarazu Collage

3. Random Forest[4]

Random Forest は集団学習法の一つである．ランダムに作成した決定木を大量に用意して予測を行う．

3.1 ブートストラップサンプル

ブートストラップサンプルは，n 組の学習データから n 組のサンプルを復元抽出したものである．ブートストラップサンプルで作成された新しい学習データの集合には，同じデータが複数存在することがほとんどである．

3.2 Random Forest の学習

Random Forest ではランダムに木を作成するが，まったくのランダムではない．次の 2 つのような手法を用いてランダム化を行っている．

1. 決定木を学習するための学習データをブートストラップサンプルとする．
2. 決定木学習中の各ノードにおいて選択候補となる属性番号の集合をランダムに選択する．

このようにして，Random Forest は学習を行っている．

4. 実験方法

4.1 学習データ

学習データは，木更津高専情報工学科 2 年生の授業プログラミング言語の試験中のコマンド履歴を用いた．コマンドは Enter を押すと同時に記録される．データは 41 名分であり，試験開始から試験終了までのデータを扱うこととした．

4.2 Random Forest を用いた実験

取得したコマンド履歴において，"emacs"，"cd"，"cp"，"gcc"，"cc"，"ls"，"./"，"send2all.sh"の数をカウントした．特徴ベクトルをそれぞれのコマンドの数およびコマンドの総使用数で構成し，Random Forest でそれらを学習する．また，目的変数は試験成績とし，試験成績をプログラミングスキルとみなした．

Random Forest の学習に用いるトレーニングデータを 31 個ランダムに選択し，残りのデータを誤差の判別に扱うテストデータとした．Random Forest を乱数の種を変更して 10 回試行し，スキルの推定精度を求めた．

5. 実験結果

Random Forest で 10 回試行をした結果，8 回目の実験で最も誤差が小さくなった．8 回目の実験での真のスキルと Random Forest による推定のスキルの結果を図 3 に示す．この時の平均誤差は 3.54，標準偏差 2.55 であった．

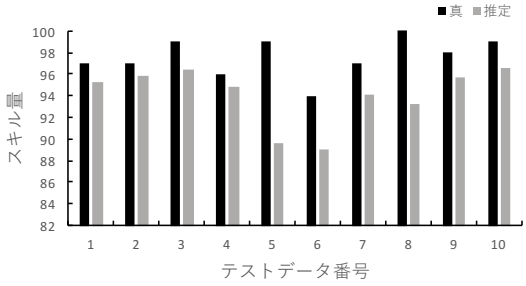

図 3　8 回目の実験での真のスキルと推定のスキル．

6. 考察

図 3 から，Random Forest による推定のスキルは，真のスキルと比較し全体的に低くなっていることがわかる．また，8 回目以外の実験では誤差が非常に大きくなってしまったものがあった．

これらの原因として，本研究で用いた学生 41 名の平均スキルが高かったことが考えられる．Random Forest では，学習時に目的変数の平均をとる．8 回目の実験ではトレーニングデータにスキルの低い学生を含んでいたため，平均値が低くなり，その影響により全体的に推定値が低くなったものだと考えられる．また，8 回目以外の実験では，テストデータにスキルの低い学生が選ばれたため誤差が大きくなってしまったと考えられる．

7. まとめ

現在の IT 教育ではプログラミングスキルは完成したプログラムのみから評価されている．本研究ではコマンド履歴に着目しプログラミングスキルを推定することを試みた．推定の手法として Random Forest を用いた．実験を行った結果，Random Forest でスキルを推定することができた．

今後の課題は，推定精度を向上させることである．また，潜在スキルの低い者の推定を考えている．早期に技術習得がうまくいっていない学生を探しだすことができれば，その手助けをすることができると考える．

謝辞　本研究は JSPS 科研費 25750095 の助成を受けたものです．

参考文献

[1] 後藤正幸, 小林学, “入門パターン認識と機械学習”, コロナ社, pp.72-107, 2014.
[2] L.Breiman, J.H.Friedman, R.A.Olshen and C.J.Stone, “Classification and Regression Trees”, Wadsworth, 1984.
[3] Quinlan, “J. R. C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, 1993.
[4] L.Breiman, "Random Forests", Machine Learning, 2001.