

広島大学大学院博士課程後期

工学研究科博士論文

ランダムフォレストを用いた非線形システムの
モデリングとその応用

(Modeling Method of Nonlinear Systems Using
Random Forest and Their Applications)

システムサイバネティクス専攻

D161632 今地 大武

目次

1 序論	1
1.1 背景	1
1.2 システムの非線形性とモデリング	2
1.3 RFによる時間変化を考慮したモデリング	4
1.4 本論文の目的と各章の構成	5
2 ランダムフォレスト	7
2.1 緒言	7
2.2 分類木	8
2.2.1 学習の手順	9
2.2.2 分類の手順	12
2.3 回帰木	13
2.4 ランダムフォレスト	14
2.4.1 学習	15
2.4.2 分類・回帰の手順	17
2.4.3 重要度 [9]	18
2.5 結言	20
3 ランダムフォレストを用いた油圧ショベル掘削操作における特徴 分析法	22
3.1 緒言	22
3.2 定量化	25
3.3 定量化の検証実験	25
3.3.1 実験装置	25
3.3.2 実験手順	27
3.3.3 検証結果	28
3.4 状態遷移モデル表現	32
3.5 状態遷移モデルの検証実験	33
3.5.1 実験手順	34
3.6 検証結果	35
3.6.1 取得データ	35
3.6.2 ランダムフォレストの学習	36
3.7 操作の特徴分析	38

3.7.1	行動モデルに基づく特徴分析	39
3.7.2	状態遷移モデルに基づく特徴分析	43
3.8	結言	48
4	ランダムフォレストによる変数選択に基づくデータベース駆動型 モデリングによる線形空燃比センサ出力予測	50
4.1	緒言	50
4.2	ランダムフォレストの重要度に基づく変数選択法	53
4.2.1	変数選択法の概要	53
4.2.2	他の変数と高い相関をもつ変数の削除	54
4.2.3	重要度に基づく非線形性を考慮した変数選択	56
4.3	データベース駆動型モデリング (DDM)	57
4.3.1	対象システムの記述	58
4.3.2	システムパラメータの同定	59
4.3.3	データベース構築	60
4.3.4	システムパラメータ同定 (オンライン同定)	65
4.3.5	数値例	65
4.4	線形空燃比センサ出力予測による検証	72
4.4.1	線形空燃比センサ出力予測	72
4.4.2	検証結果	73
4.5	結言	75
5	結言	77
	参考文献	80

1 序論

1.1 背景

Industry 4.0 や Society 5.0 に代表されるように，IoT や AI 技術を応用して，従来までの人がシステムに適応する構造を，システムが人に適応する構造に変化させる動きが活発になっている。これらを実現するためには，周りの状況に合わせてシステムの挙動を自律的に改善する仕組み，すなわち自動制御技術を中心とした新たなプラットフォームの構築が必要である。このような状況を受けて，現在の制御技術の分野では，システムモデリング，データの抽出・解析，コントローラ的设计等を担うプラットフォームであるスマートプラットフォームの構築など，高度な制御技術の実現を目指して様々な研究が進められている [1,2]。その概要図を図 1.1 に示す。スマートプラットフォームは，システムモデル，入出力データ，制御結果等，様々な情報を保持しており，制御対象に応じて，必要な要素を必要なだけ提供し，制御系を構築する。このプラットフォームを実現するためには，制御対象の特徴を捉え，解析する技術が必須となる。したがって，本研究では非線形システムの様々な特性に，構造解釈が容易なモデリングの応用を通して対応する手法について考察

する。

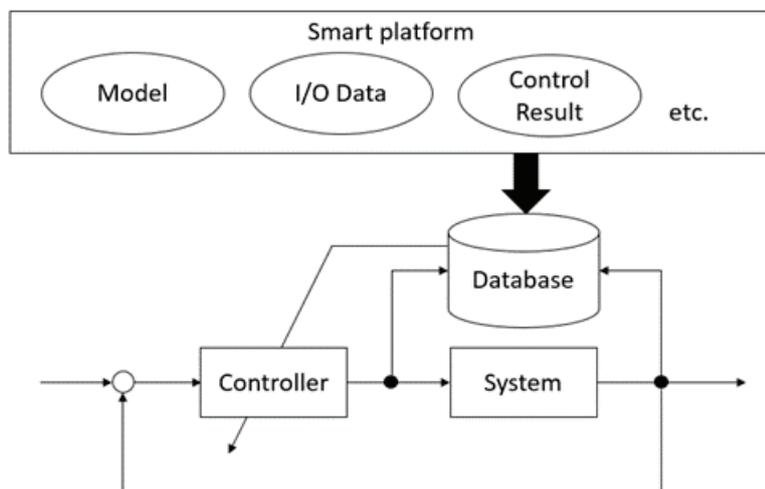


図 1.1: スマートプラットフォーム概要図

1.2 システムの非線形性とモデリング

産業界のシステムの多くは多入力であり、さらに非線形かつ動的な特性、および時間変化する内部状態を有しているため、モデリング技術には、これらの特徴を考慮することが求められている。モデリング技術はホワイトボックスモデリングとブラックボックスモデリングに大別される。前者はシステムの物理的特性から、数式によってモデルを構築する手法である。一方、後者は、システムの構造が不明であり、入出力信号のみが与えられている場合に、入出力関係に着目し汎用的なモデルを構築する手法である。一般的に、産業システムは複雑な数学的構造を有

しているため、ホワイトボックスモデリングの適用は困難である。したがって、本論文ではブラックボックスモデリングを取り扱う。

ブラックボックスモデリングを実現する手法として、機械学習技術を用いる方法がある。この方法は、システムの入出力関係をヒューリスティックな手法に基づいてモデリングする方法である。機械学習を用いた代表的な手法として、ディープラーニングを用いる手法が挙げられる [3]。ディープラーニングは、ニューラルネットワークの一種であり、複数の事前学習器を重ねた構造をもつ。これにより、高レベルの学習が実現できるため、複雑なシステムのモデリングが可能となる。一方で、設計が難しく、かつモデルの構造解釈が極めて困難であるため、ディープラーニングは、画像処理のように内部構造の解析が必須ではないシステムのモデリングには有効ではあるが、本論文で取り扱う産業システムのように、安全面考慮やインシデント対応次第で生成されたモデルの解析が必要になる場合には不向きであると考えられる。

一方、比較的構造解釈の容易な機械学習手法として、決定木が提案されている [4]。決定木はその構造の解釈の容易さから、モデリングのほかにもシステムの解析に用いられる [5]。しかしながら、過学習を起こしやすく、外れ値やノイズの多いデータを用いたモデリングには不向きであると考えられる。ところで、ランダムフォレスト (Random Forest: RF) と呼ばれる複数の決定木を統合することで過学習を抑制する手法が提案されている [6]。この手法は、先述のディープラーニングに比べると、一般的に構造解釈が容易な一方で、モデルの表現能力は劣るが、本研究で対象とする産業システムへの適用としては、十分な性能を発揮できると

考えられる。したがって、本手法の特性を利用することにより、非線形性の考慮と、構造解釈の容易さ確保の両立を目指す。

1.3 RF による時間変化を考慮したモデリング

先述のとおり、RF は非線形性と構造解釈の容易さの両立が可能な一方で、システムの時間変化する特性を取り扱うことが難しい。これは、RF が自己回帰項などの過去のシステムの情報を取り入れることが難しいためである。したがって、本論文では、時間変化する特性を取り入れることが可能な手法に対して、RF により解析機能を付与することで、時間変化する特性の解析手法の確立を目指す。本論文では、非線形システムの動的特性を考慮したモデリングと解析を実現するために、2種類の手法を考える。一つは、まずシステムの状態を複数定義し、システムの時間変化を状態遷移モデルとして表すことで、時間変化する特性をRFによって解析・モデリングする方法である。具体的には、RFの内部構造を利用してシステムの状態を定量化し、その結果から状態遷移時に表れる特徴を解析する手法である。本手法は油圧シャベルによる土砂のダンプ積み込み作業データにより実証実験を行う。

もう一つは、RFとシステムの動的特性を考慮可能な手法を組み合わせる手法である。本手法によってシステムのモデリングに必要なデータを解析し、その結果をモデリングに利用する。本手法は自動車の線形空燃比センサの出力予測により実証実験を行う。

1.4 本論文の目的と各章の構成

本論文では、動的な特性を考慮でき、かつ解析への応用が容易な非線形システムモデリングの実現を目的に、先述の RF を中心とした 2 つの手法について考察する。

まず、第 2 章では、本論文の中心となる技術である RF について述べる。RF は機械学習の一種であり、解析が比較的容易な決定木と呼ばれる学習機によって構成されているため、構造解釈の容易さを前提としたモデリングが可能である。前半で決定木の構造とアルゴリズムについて述べたのち、RF について解説する。さらに、RF を用いた説明変数の重要度算出についても述べる。

第 3 章では、RF による状態の定量化と、状態遷移モデルを用いた解析手法について述べる。一般的に、RF は時系列データを取り扱うことが難しいため、RF の構造を状態遷移モデルに落とし込むことによって、RF による時系列データの解析を試みる。本章の後半は、油圧ショベルの掘削作業の解析を通して、本手法の有効性を検証する。作業員の操縦する油圧ショベルを解析対象のシステムとして、各作業工程をシステムの状態とみなすことで本手法を適用し、作業員のスキルを定量的に評価する。

第 4 章では、RF による解析に加えて、時系列データのモデリングについて述べる。動的特性に特徴のあるシステムに対して、システムの動的特性と非線形性を考慮可能なデータベース駆動型モデリング (DDM) に、RF の解析機能を搭載することにより、解析とモデリングを同時に行える手法を提案する。本手法により、モデルの精度向上が期待できる。なお、本手法は自動車の線形空燃比センサ (LAFS) の出力予測に適用すること

により、有効性について検証する。LAFSはエンジンの燃焼を制御する上で重要な役割を担っているが、自動車の走行状態や経年劣化によって動的特性が変動する。したがって、本章ではRFによって改良したDDMを用いて、動的特性に影響を与えている要因の特定、動的特性の推定、本来のLAFS値の推定を行う。

第5章では本研究の総括をするとともに、残された課題点について言及し、今後の解析可能なモデリングの展望について述べる。

2 ランダムフォレスト

2.1 緒言

本章では、本論文で用いられる基礎技術である、ランダムフォレスト (RandomForest:RF) について述べる。RF とは決定木 [4] を構成要素とした機械学習技術の一つであり、2002年にLeo Breimanにより提案された [6]。RFは、集団学習と呼ばれる手法の一種であり、従来の機械学習手法に比べ、多くの計算リソースを要する一方で、予測誤差の低減が期待できる。機械学習における予測誤差は、主に未学習か過学習により引き起こされる。未学習とは、学習データに対する学習が十分に行われていない状態であり、過学習とは、過剰に学習をしている状態である。未学習の場合、モデリング対象に対して、十分なモデルが生成されていないため、システムの一部の特性を表現できず、予測誤差を生じる。過学習の場合、学習データを過剰に学習することにより、モデリングの対象外である計測誤差などのノイズを学習することで予測誤差を生じる。一般的に未学習と過学習はトレードオフ関係にあるため、モデリングをする際には、対象に合わせてパラメータ設計を行う必要がある。

このような状況に対し、近年、集団学習が提案されている。集団学習

とは、決定木などの単純な構造を持つ学習器（弱学習器）を複数束ねた構造をもつ機械学習手法である。集団学習には、弱学習器を並列に学習するバギング [7] と、逐次的に学習するブースティング [8] がある。RF は、バギングの一種であり、弱学習器として決定木を採用している。バギングは、並列に学習した複数の弱学習器を統合することで、過学習が元となる予測誤差を低減することができる。このとき、過学習による予測誤差は、弱学習器の多さ、つまり計算リソースの増大とトレードオフとなる。したがって、計算リソースを用意することで、過学習と未学習のトレードオフとは別に、過学習の改善が可能となる。RF は、さらに変数のサンプリングを行うことにより、変数の予測精度への貢献度合いを重要度として評価することができる。

決定木とは、葉と呼ばれるノードと、枝と呼ばれる分岐によって構成される機械学習手法の一手法である。一般的な機械学習同様、分類と回帰の両方が可能であるが、それぞれに合わせて構造を若干変える必要がある。分類、回帰を行う決定木のことをそれぞれ分類木、回帰木と呼ぶ。以降では、まず RF の弱学習器である決定木について述べ、次に RF について述べる。

2.2 分類木

分類木の学習について説明する。分類木の学習に使用するデータは、特徴ベクトルとラベルである。特徴ベクトルは、分類するデータの特徴を保持する 1 次元ベクトル x_i の集合 X を示す。式 (1) に特徴ベクトルを示す。次に、ラベルは、分類先を表す自然数 y_i の集合 Y である。式 (3) に

特徴ベクトルを示す。

$$X = \{\mathbf{x}_i | i = 1, 2, \dots, n_1\} \quad (1)$$

$$\mathbf{x}_i = \{x_{ij} | j = 1, 2, \dots, n_2\} \quad (2)$$

$$Y = \{y_i | i = 1, 2, 3, \dots, n_1\} \quad (3)$$

$$(y_i = 1, 2, 3, \dots, K) \quad (4)$$

ここで n_1 は訓練データの数, n_2 は特徴ベクトルの次元数, K はクラス数, i は訓練データの通し番号である。 \mathbf{x}_i , y_i はそれぞれ特徴ベクトルとラベルであり, x_{ij} は \mathbf{x}_i の j 番目の要素である。

2.2.1 学習の手順

特徴ベクトルとラベルのペアで構成される訓練データの集合を T とする。

$$T = \{t_i | i = 1, 2, 3, \dots, n_1\}, \quad (5)$$

$$t_i = \{\mathbf{x}_i, y_i\} \quad (6)$$

また, 分類木の深さを d とし, 添え字 r か l を表す $c_k (k < d)$ を定義する。以下に学習手順を Step ごとに示す。

STEP 1 初期化

$d = 0$ を設定する。

STEP 3 変数候補の決定

過学習を抑制するため、変数のランダムサンプリングを行う。ランダムサンプリングされた説明変数番号の集合 I_j を以下のように定義する。

$$I_j = \subset \{1, 2, 3, \dots, n_2\} \quad (7)$$

ここで、 I_j は $1, 2, 3, \dots, n_2$ の中から重複を許さず無作為に選んだ $m(\leq n_2)$ 個の自然数の集合である。

STEP 3 分岐条件の決定

分岐条件は、説明変数の番号 \bar{j} と、説明変数の閾値 a によって定義する。詳細は式 (9) の通りである。

$$\bar{j} \in I_j \quad (8)$$

$$\min_k(x_{k\bar{j}}) \leq a \leq \max_k(x_{k\bar{j}}) \quad (9)$$

$T_{c_1 c_2 \dots c_d}$ のうち、 $x_{i\bar{j}} \leq a$ であるような t_i の集合を $T_{c_1 c_2 \dots c_d 1}$ 、 $x_{i\bar{j}} > a$ であるような t_i の集合を $T_{c_1 c_2 \dots c_d 2}$ とする。 \bar{j} 、 a はデータが可能な限りラベル別に分かれるように設定する。したがって、以下の式 (10) を最小にするような \bar{j} と a を求め、それぞれ \hat{j} 、 \hat{a} する。

$$J = I(T_{c_1 c_2 \dots c_d}) - \sum_{k=1}^2 P(T_{c_1 c_2 \dots c_d k} | T_{c_1 c_2 \dots c_d}) I(T_{c_1 c_2 \dots c_d k}) \quad (10)$$

$$I(T_{c_1 c_2 \dots c_d}) = 1 - \sum_{k=1}^K P^2(C_k | T_{c_1 c_2 \dots c_d}) \quad (11)$$

$P(T_{c_1 c_2 \dots c_d k} | T_{c_1 c_2 \dots c_d})$ は $T_{c_1 c_2 \dots c_d}$ に占める $T_{c_1 c_2 \dots c_d k}$ の要素の割合である。 C_k は、 $y_i = k$ である t_i の集合であり、 $P^2(C_k | T_{c_1 c_2 \dots c_d})$ は $T_{c_1 c_2 \dots c_d}$ に占める C_k の要素の割合である。式(10)はジニ係数であり、格差が大きいほど値が小さくなる統計量である。本稿ではデータの分離度合いを、ラベル間のデータ数の格差をジニ係数で表している。

STEP 4 分割

Step2で決めた \hat{j} , \hat{a} に従い、 $T_{c_1 c_2 \dots c_d}$ を $\hat{T}_{c_1 c_2 \dots c_d 1}$ と $\hat{T}_{c_1 c_2 \dots c_d 2}$ の2つの集合に分割する。また、 \hat{j} , \hat{c} と、 $T_{c_1 c_2 \dots c_d}$ のラベル情報を $B_{c_1 c_2 \dots c_d}$ に保存する。 $T_{c_1 c_2 \dots c_d}$ のラベル情報を $Y_{c_1 c_2 \dots c_d}$ とすると、 $B_{c_1 c_2 \dots c_d}$ は以下で表される。

$$B_{c_1 c_2 \dots c_d} = \{\hat{j}, \hat{c}, Y_{c_1 c_2 \dots c_d}\} \quad (12)$$

STEP 5 終了判定

$\hat{T}_{c_1 c_2 \dots c_d}$ が以下に示す終了条件を満たさない場合、 $d \rightarrow d + 1$ とし、 $c_d = 1$ として、Step2の操作を行う。その後、 \hat{T}_2 も同様に終了条件を満たさない場合、 $d \rightarrow d + 1$ とし、 $c_d = 2$ として、Step2の操作を行う。

終了条件は以下の通りである。

- ① データの集合が、単一のラベルのデータのみ
- ② データの集合内のデータの数、一定数以下

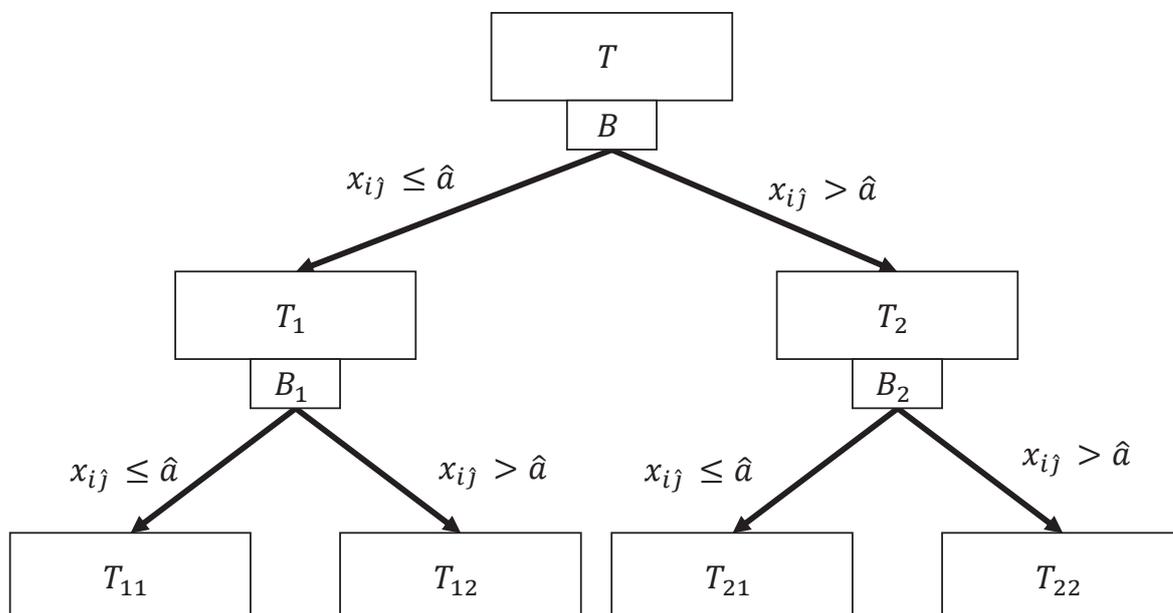


図 2.1: 決定木の概要図

③ 所定の分割数に到達

以上が、分類における決定木の学習手順であり、 $B_{c_1 c_2 \dots c_d}$ が学習結果となる。

2.2.2 分類の手順

決定木の分類について説明する。分類対象の特徴ベクトルを $\bar{x} = \{\bar{x}_j | j = 1, 2, \dots, m\}$, 決定木の深さを d とする

以下に決定木の分類手順を示す。

STEP 1 初期化

$d = 0$ とする。

STEP 2 条件の有無の確認

$B_{c_1c_2\cdots c_{d1}}$ および $B_{c_1c_2\cdots c_{d2}}$ が存在する場合は Step3 へ進み、存在しない場合は Step4 へ進む。

STEP 3 ルール適用

$B_{c_1c_2\cdots c_{d1}}$ の \hat{j} , \hat{c} に関して, $\bar{x}_{\hat{j}} \leq \hat{c}$ ならば $d \rightarrow d + 1$, $c_d = 1$ とし, $\bar{x}_{\hat{j}} > \hat{c}$ ならば $d \rightarrow d + 1$, $c_d = 2$ として Step2 を行う。

STEP 4 結果の出力

$B_{c_1c_2\cdots c_d}$ 中の $Y_{c_1c_2\cdots c_d}$ で一番データ数の多いラベルを, 分類先 \hat{y} として出力する。

2.3 回帰木

回帰手法としての決定木の学習について説明する。決定木の学習に使用するデータは, 特徴ベクトルと目的変数である。特徴ベクトルは式 (1) と同様だが, 目的変数は式 (3) とは異なり連続値となる。目的変数 Y を式 (13) に示す。

$$Y = \{y_i | i = 1, 2, 3, \dots, n_1\} \quad (13)$$

ここで、 y_i は実数である。回帰の場合、分岐条件の決定に式 (10) を用いることができないため、代わりに以下の式を最小化する。

$$J = \sum_{k=1}^2 V(T_{c_1 c_2 \dots c_d k}) \quad (14)$$

$V(\cdot)$ は目的変数の分散である。式 (14) は、分割後の分散の和である。

終了条件も下記の通り変更する。

- ① データの説明変数の分散が一定値以下
- ② データの集合内のデータの数在一定数以下
- ③ 所定の分割数に到達

2.4 ランダムフォレスト

決定木は生成されるモデルの構造がルールの集合であるため、学習・解釈が容易である一方で、過学習を起こしやすいという問題点がある。一般的な決定木の過学習対策としては、主に下記の3つがある。

- ① 木の深さを制限する。
- ② 葉のデータ数を一定以上に制限する。
- ③ 過学習をさせたのち、枝を刈る。

1, 2 は一般的な機械学習手法と同様、学習を打ち切る条件を設定する手法である。3 は決定木の構造を利用した手法であり、一旦過学習を

許しながら、決定木を学習する。その後、予測精度等を参考にしながら、末端から順番に枝を刈っていくことで、過学習部分の構造の除去を目指す。この手法は式(11)(14)に示す評価規範の最小化にあまり貢献していないルールへの対策として有効な一方で、図2.2のように却って予測精度を下げる場合がある。したがって、本稿では集団学習の1つであるランダムフォレスト(Random Forest:RF)を採用する。

RFは複数の決定木で構成されており、それぞれの決定木を学習させる必要がある。決定木同様、分類と回帰の両方が実現可能であり、それぞれアルゴリズムがわずかに異なる。

2.4.1 学習

学習アルゴリズムは分類木、回帰木で共通となる。以下に、本稿でのRFの学習アルゴリズムを示す。

STEP 1 ブートストラップサンプリング

学習の事前準備として、RFの訓練データから決定木の数だけ訓練データを生成する。決定木の数 n_3 とする。RFの訓練データを式(1)(3)と同じように考えると、決定木の訓練データは以下となる。

$$X_l = \{\hat{\mathbf{x}}_i | i \in I_l\} \quad (15)$$

$$Y_l = \{\hat{y}_i | i \in I_l\} \quad (16)$$

$$l = 1, 2, 3, \dots, n_3 \quad (17)$$

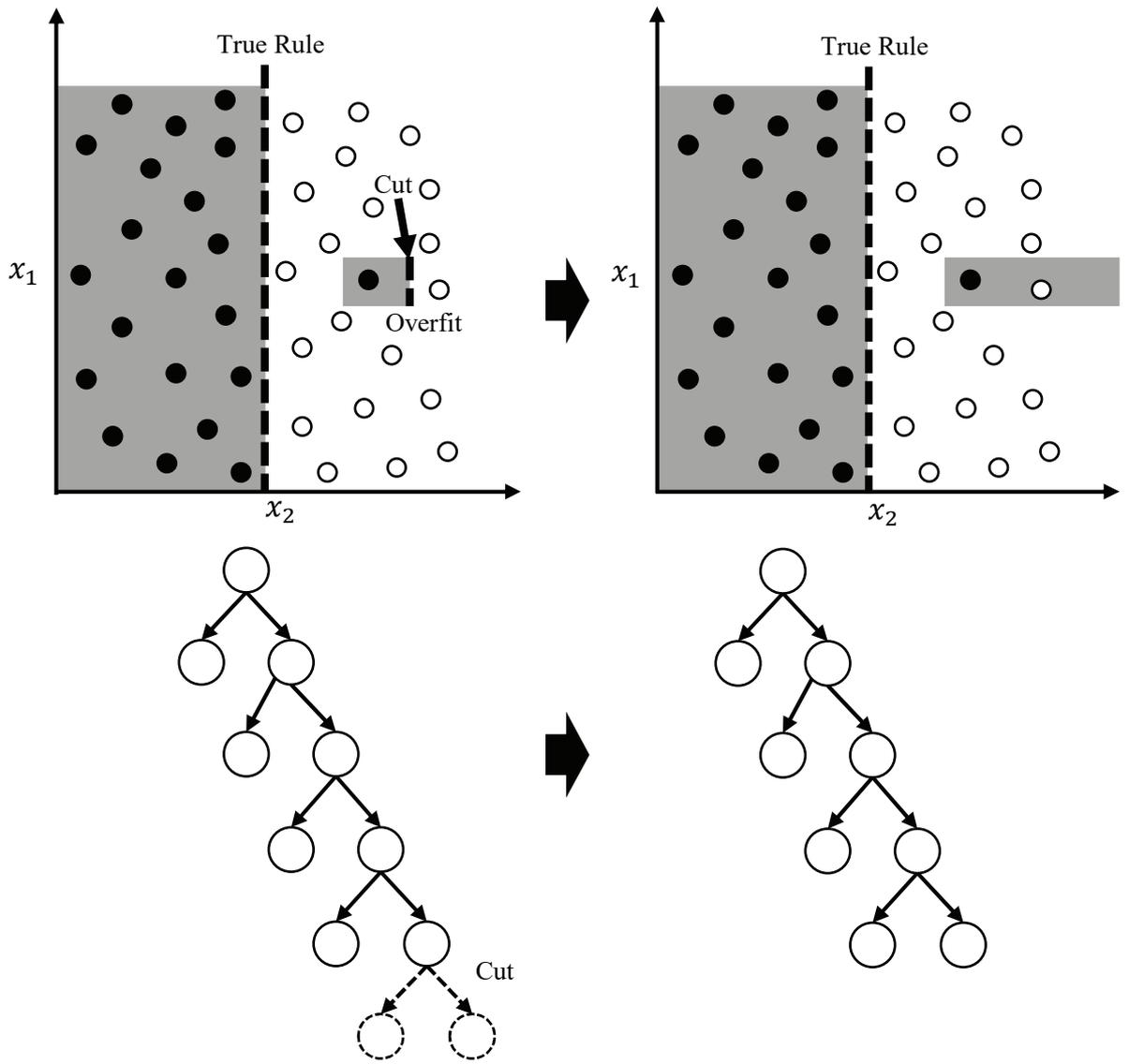


図 2.2: 枝刈りの問題点

ここで、 I_l は $1, 2, 3, \dots, n$ の中から重複を許してランダムに選んだ $\hat{n} (\leq n)$ 個の自然数の集合である。式 (15)–(16) のようなデータセットを決定木の数だけ生成する。以上の操作をブートストラップサンプリングと呼ぶ。

STEP 2 決定木の学習

学習に使用する。 X_l, Y_l を前節の X, Y とし、それぞれの決定木を学習する。

ブートストラップサンプリングで使用されなかったデータの集合を Out of Bag (OOB) と呼び、後述する重要度の算出に用いる。OOB を式 (18)(19) に示す。

$$X_{oob} \not\subset X_1 \cup X_1 \cup \dots \cup X_{n_3} \quad (18)$$

$$Y_{oob} \not\subset Y_1 \cup Y_1 \cup \dots \cup Y_{n_3} \quad (19)$$

ここで、 X_{oob} は OOB の説明変数、 Y_{oob} は OOB のラベルデータおよび目的変数を表している。

2.4.2 分類・回帰の手順

分類木を採用した場合、RF が分類するクラスは分類木の多数決で決定する。一方で回帰木を採用した場合は、回帰木の出力の平均を RF の出力とする。以上の操作により、過学習を起こした決定木がある場合でも、その影響を最小限にとどめることができる。詳細を図 2.3 に示す。

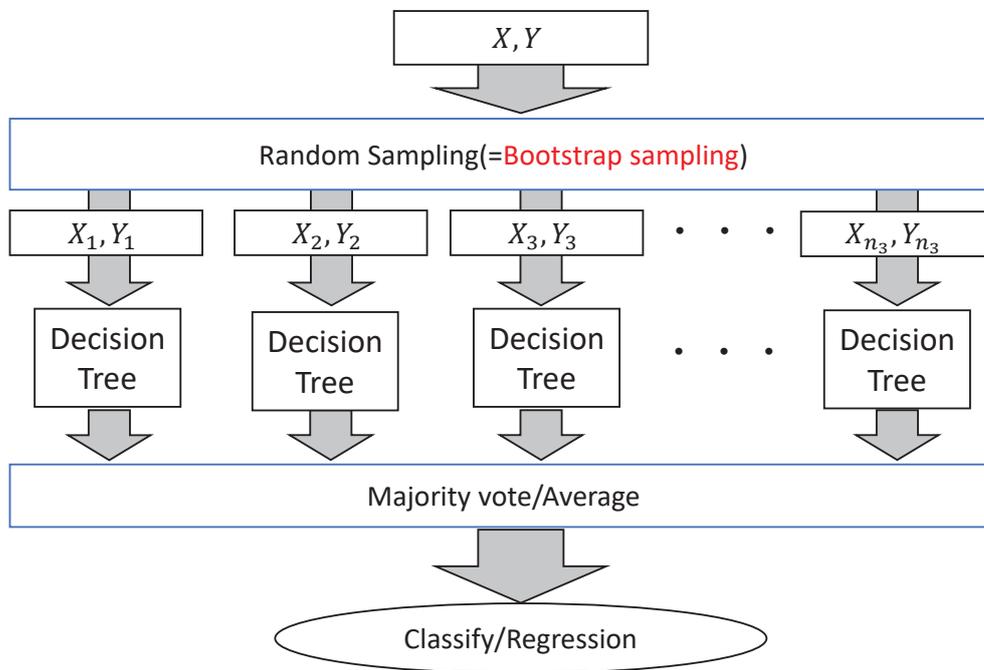


図 2.3: ランダムフォレスト概要図

2.4.3 重要度 [9]

RFにおける重要度とは、各変数がどの程度分類・回帰精度に貢献をしているか定量化した数値であり、前述の OOB を基に算出する。以下に手順を示す。

STEP 1

OOB を用いて予測し、予測誤差を算出する。分類の場合は分類の不正回数、回帰の場合は式 (20) に示す誤差を予測誤差として算出

する。

$$J_{oob} = \frac{1}{n_{oob}} \sum_{k=1}^{n_{oob}} (\hat{y}_k - y_k)^2 \quad (20)$$

ここで、 J_{oob} は予測誤差、 n_{oob} は OOB 内のデータ数、 \hat{y}_k は説明変数の予測値、 y_k は説明変数の真値である。

STEP 2

評価対象の変数の値を、OOB 内で無作為に入れ替えたものを、新たなデータセットとする。

STEP 3

新たなデータセットに対して STEP 1 を実行し、予測誤差を算出する。

STEP 4

STEP 3 の結果から、STEP 1 の結果を減算した結果を変数の重要度として出力する。

予測結果に関与しない変数は、OOB 内で無作為に値を入れ替えても予測誤差が増えないため、STEP 4 の計算結果は小さくなる。一方で、予測結果に関与する変数は、値を無作為に入れ替えると予測誤差が増大するため、STEP 4 の計算結果は大きくなる。したがって、STEP 4 で算出される重要度は、予測精度に対する変数の影響の大きさを表しているといえる。

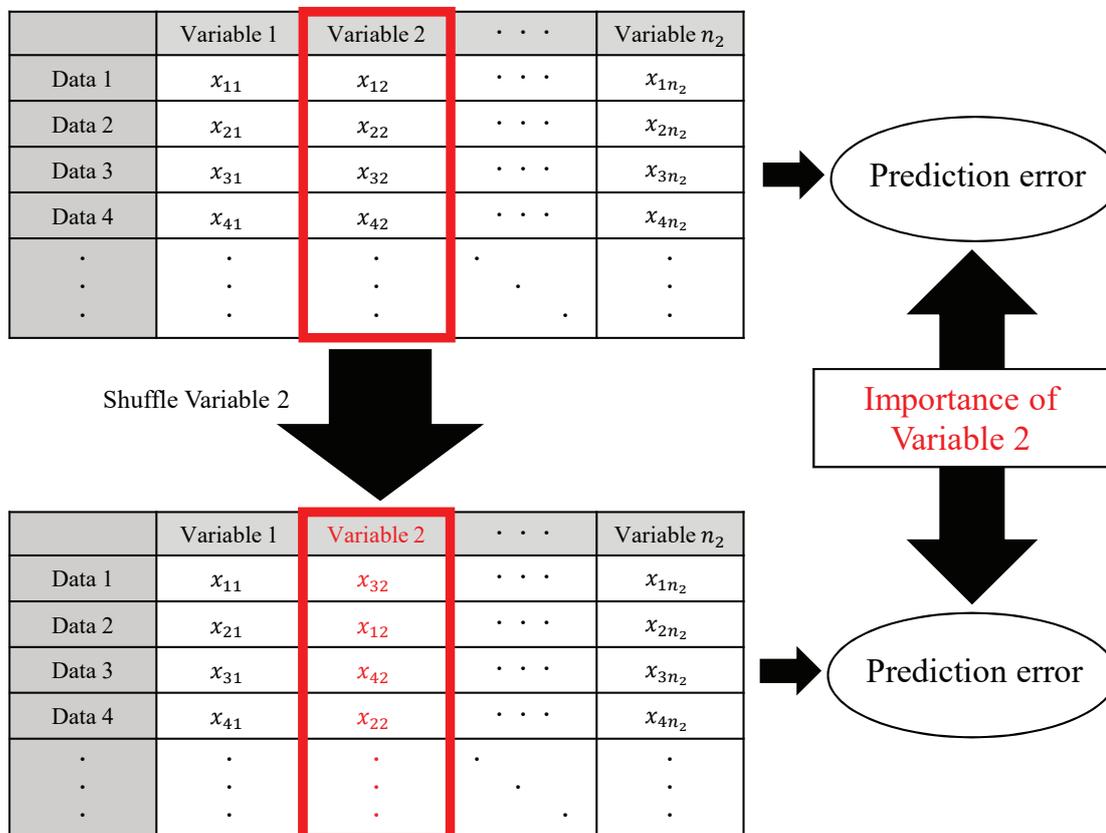


図 2.4: 重要度の算出概要図

2.5 結言

本章では、提案手法の中心技術であるランダムフォレスト (Random-Forest:RF) について述べた。RF は複数の決定木で構成されており、解析性と非線形性の考慮を両立した手法である。さらに、説明変数の予測精度への貢献度を重要度として定量化することが可能であり、後述する提案手法において、モデリングおよび解析に活用している。そこで本論文

では、この RF を用いたモデリング・解析技術を提案し、それぞれ異なる実機にて有効性を検証する。第 2 章では、まず RF が決定木によって構成されていることを利用した解析手法を提案し、油圧ショベルの作業者スキルの解析を通じて有効性を検証する。第 3 章では重要度を用いた解析とモデリング手法を提案し、線形空燃比センサの出力予測によって有効性を検証する。

3 ランダムフォレストを用いた油圧ショベル掘削 操作における特徴分析法

3.1 緒言

第2章で述べたとおり，RFは個々の決定木の出力の多数決や平均をとることにより，決定木の過学習の影響を最小限にとどめている。言い換えれば，RF全体として過学習が抑えられている場合でも，構成要素である決定木は過学習により誤判定を行っている可能性がある。本章では，この特徴を利用してシステムの状態解析を行う。具体的には，状態解析を行う上で見本となるシステムのデータを用いてRFを訓練し，生成されたモデルに解析対象データを入力することで，見本システムと解析対象の差を決定木のクラスの割合として定量化する。解析対象の説明変数の中に見本システムにとってのノイズや外れ値が多ければ，それに合わせて決定木の誤判定の割合も大きくなる。したがって，本手法を用いることで，見本システムと解析対象の差を定量的に表すことができると考えられる。概要を図3.1に示す。本章では油圧ショベルの作業スキル定量化を通して，本手法の有効性を検証する。ここで，本検証における見本

システムは、油圧ショベル操作の熟練者とする。

建設現場では、生産性向上を課題として国土交通省主導のもと、i-Construction と呼ばれる取り組みが進められている [10]。i-Construction は、「ICT（情報通信技術）の全面的な活用」を主眼に、3D データの活用、さらには IoT（Internet of Things）や AI（Artificial Intelligence）の革新的な技術を施工現場に導入することで、建設生産システム全体の生産性向上を図り、魅力ある建設現場を目指すこととしている [11]。これにより、建設機械の自動運転や施工管理システムのデータ化・自動化により効率化が進められている。[12-14]

現在の建設現場では、効率的な作業を行うために建設機械が欠かせない。建設機械は、その操縦技術に個人差があり、作業効率は作業者に依存する。すなわち、熟練者による最適化された建設機械の操縦技術により、建設現場の効率的な作業が実現している [15]。そのため、作業者の作業テクニックやコツといった特徴を定量的に評価し、効率的な作業を把

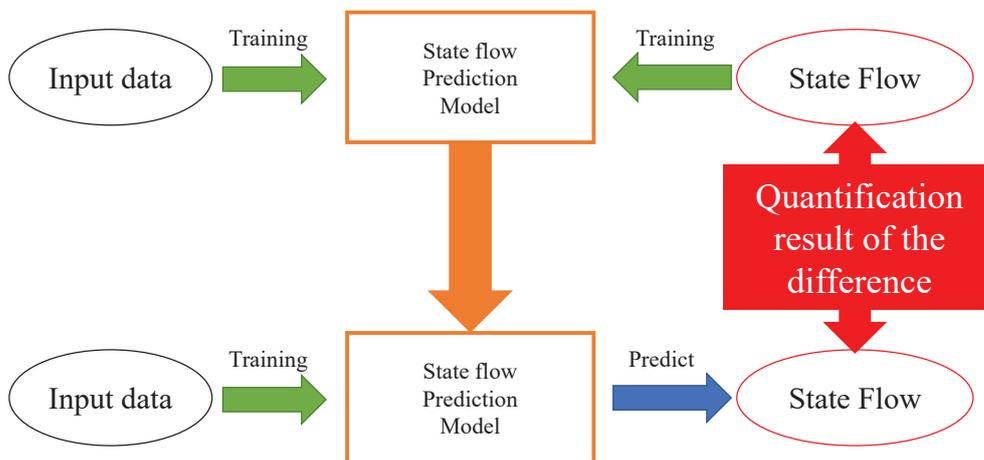


図 3.1: 定量化概要図

握することで、建設現場をさらに効率的に稼働させることができる。作業機械の操作における技量やその特徴を抽出する先行研究として、視覚情報に関する認知能力についての技量抽出 [16]、機体操作時の挙動の差異に着目した技量抽出 [17]、制御工学的アプローチに基づく操作スキルの評価 [18]、バケット爪先軌跡の差異による評価 [19] 等が実施されている。これらの研究で、熟練作業者は、操作パターンや爪先軌跡、さらにはゲイン変化にばらつきが少なく、一定した操作ができる特徴を有していることが述べられている。このように熟練作業者は規則的なパターンで操作することが可能であり、一方で、非熟練作業者であるほど、そのパターンにばらつきがあることが知られている。しかし、これらの手法は主に定性的な評価であり、作業の効率化を前提とした操作スキルの評価手法としては課題が残る。

本章では、まず繰り返し精度が良い熟練作業者の操作パターンを基準として、一般作業者と非熟練作業者の作業分類を実施し、その分類結果を定量的に評価する。さらに、油圧ショベルの操作パターンを機械学習で学習させ、学習結果に基づき、状態遷移モデルで表現することにより、操作の特徴を分析する手法を検証する。具体的には、はじめに、油圧ショベルの熟練作業者の操作データを基準の操作パターンとするため、熟練作業者の代表的な操作パターンを抽出し、操作量を入力、作業状態を出力として RF に学習させる。ここで、熟練作業者は、繰り返し操作の精度が良い特定の作業者を選定する。次に、熟練作業者、一般的な作業者、非熟練作業者の操作量を、学習した RF にそれぞれ入力し、作業状態の判別を行う。さらに、作業判別の結果をそれぞれ状態遷移モデルで表現

し、その差異を考察することで操作の特徴を分析する。

3.2 定量化

RFに含まれる決定木の数を N とする。各決定木から出力された \hat{y} のうち、値が $i (i = 1, 2, 3, \dots, K)$ であるものの総数を l_i とする。このとき、以下のような C_i を定義する。

$$C_i = \frac{l_i}{N} \quad (21)$$

以上に定義される C_i を定量化の結果とする。本研究では、 i はそれぞれ1つの作業状態に対応しており、特に「待機」状態に着目した定量化を行う。

3.3 定量化の検証実験

3.3.1 実験装置

操作スキルの定量化に用いる油圧ショベルは、コベルコ建機株式会社製 SK200-9 アセラ・ジオスペックである。実験装置を図 3.2 に示す。

対象操作は、土砂のダンプ積み込み作業における操作とし、そのスキルを定量化する。一般的な油圧ショベルによる土砂のダンプ積み込み作業は「掘削 (Dig)」、「持ち上げ旋回 (Lift)」、「排土 (Dump)」、「復帰旋回 (Reposition)」の4つの状態に分類される。本検証では、これらの分類に作業終了後の無操作状態である「待機」状態を加えた5つの状態に作業

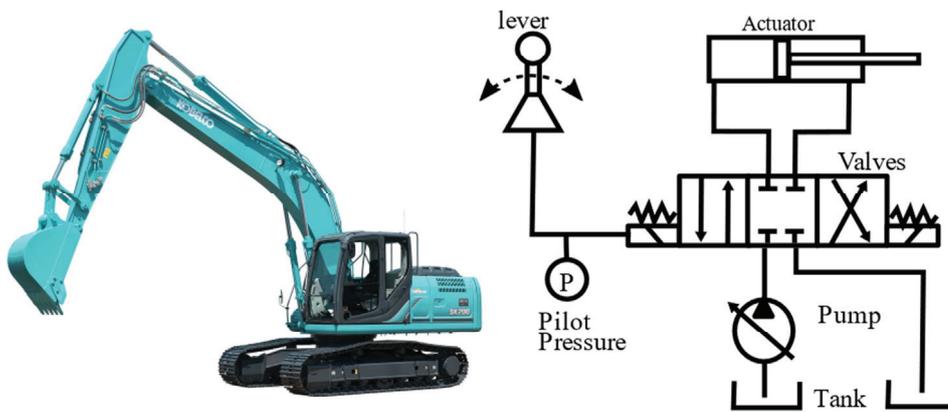


図 3.2: 油圧ショベル (SK200-9)

を分類する。作業状態の概要を図 3.3 に示す。油圧ショベルの操作は、「バケット」、「アーム」、「ブーム」、「旋回」の 4 つの操作で構成される。油圧ショベルは、バルブ内のスプールを動かすことにより、油圧ポンプから油圧アクチュエータへの流路が確保され、それぞれの油圧アクチュエータが稼働する。バルブのスプールは、レバー操作に応じて発生するパイロット圧が入力されることにより稼働する。そのため、操作入力量として各操作のパイロット圧を計測する。

RF に使用する特徴ベクトルの要素は、各操作の入れ側と戻し側の計 8 種類のパイロット圧とした。

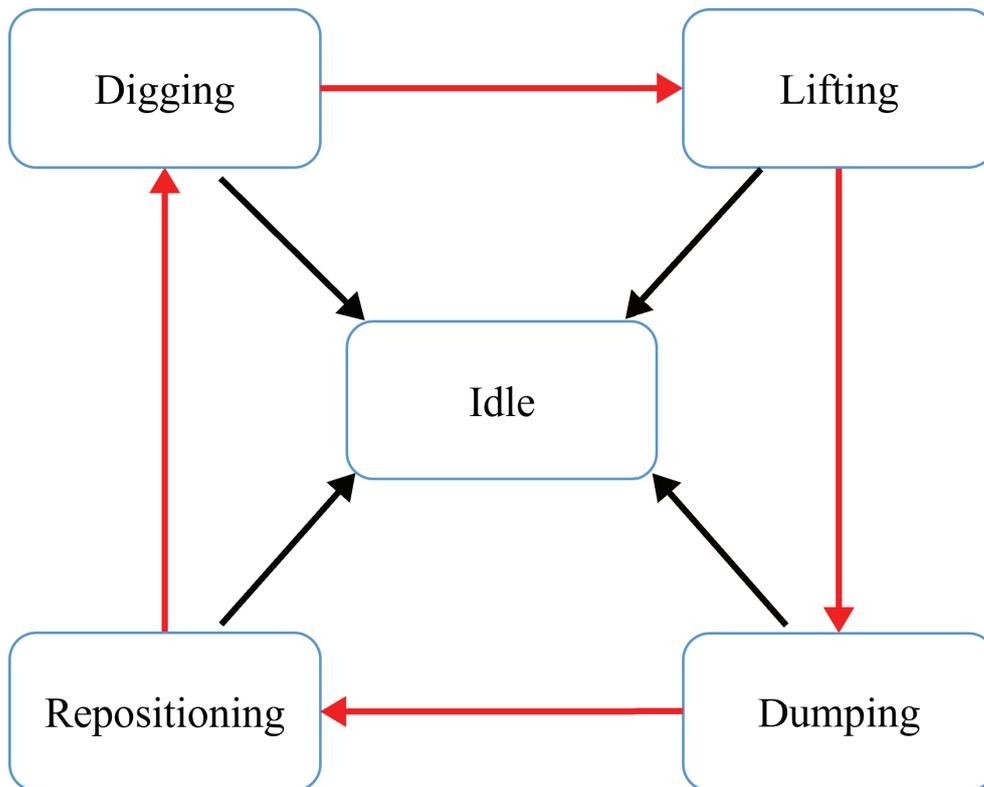


図 3.3: 土砂のダンプ積み込み作業

3.3.2 実験手順

実験手順は以下の通りである。

STEP 1 データ取得

非熟練者，一般作業者，熟練者のそれぞれの土砂のダンプ積み込み作業のパイロット圧データを取得する。

STEP 2 RF 学習

熟練者のパイロット圧データを特徴ベクトルとして，作業状態を表すラベルで RF を学習する。

STEP 3 定量化

Step1 で得られた学習済みの RF を用いて、非熟練者、一般作業員、熟練者の操作を分類し、予測ラベルの割合を出力する。予測ラベルの構成割合のうち、「待機」状態を表すラベルの割合を、操作スキルの非熟練度とする。

3.3.3 検証結果

RF の訓練データは熟練者の「待機」、「掘削」、「持ち上げ旋回」、「排土」、「復帰旋回」の合計 4309 ステップのパイロット圧データを用いた。検証用データとして熟練者、一般作業員、非熟練者のパイロット圧データをそれぞれ 4000 ステップずつ用意した。ランダムフォレストの決定木の数を 21、決定木に与えるデータの数を 1436、決定木に与えるデータの特徴量の数を 3 とした。

評価結果を図 3.4-3.6 および表 3.1 に示す。図 3.4-3.6 の Idle は「待機」状態と判定した決定木の割合 (以下、待機割合) を時系列で示したグラフである。表 3.1 はそれぞれの作業員の待機割合である。表 3.2 は、結果にランダム要素を持つ RF の有効性の検証のために、同じ訓練データと検証データを用いた RF による学習と分類を 50 回繰り返した際の、待機割合の数値検証結果である。図 3.4-3.6 と表 3.1 より、作業状態を「待機」状態と判定する決定木の割合が熟練者、一般作業員、非熟練者の順に小さいことがわかる。また、表 3.2 の検証結果より、これらの結果は有意であるといえる。

「Dump」と分類されたデータを抽出し、図 3.7-3.9 に示す。ここで、図

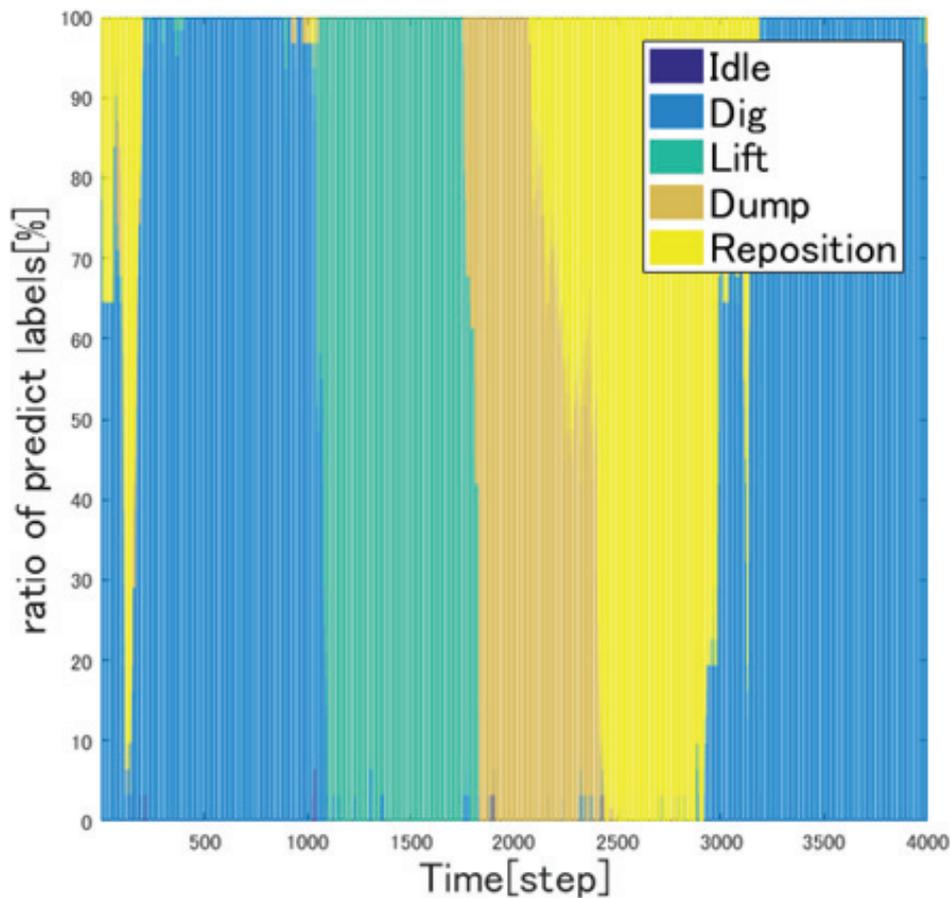


図 3.4: 熟練者の分類結果

表 3.1: 待機状態の割合

作業者	熟練者	一般	非熟練
待機割合 [%]	0.09	5.32	13.7

中の Dump はバケット排土, Arm Push はアーム押し, Swing は旋回時の各パイロット圧であり, Idle Ratio は待機割合である。この結果, 熟練者に比べて, 操作量が不足していると, 操作が「待機」状態の割合が高

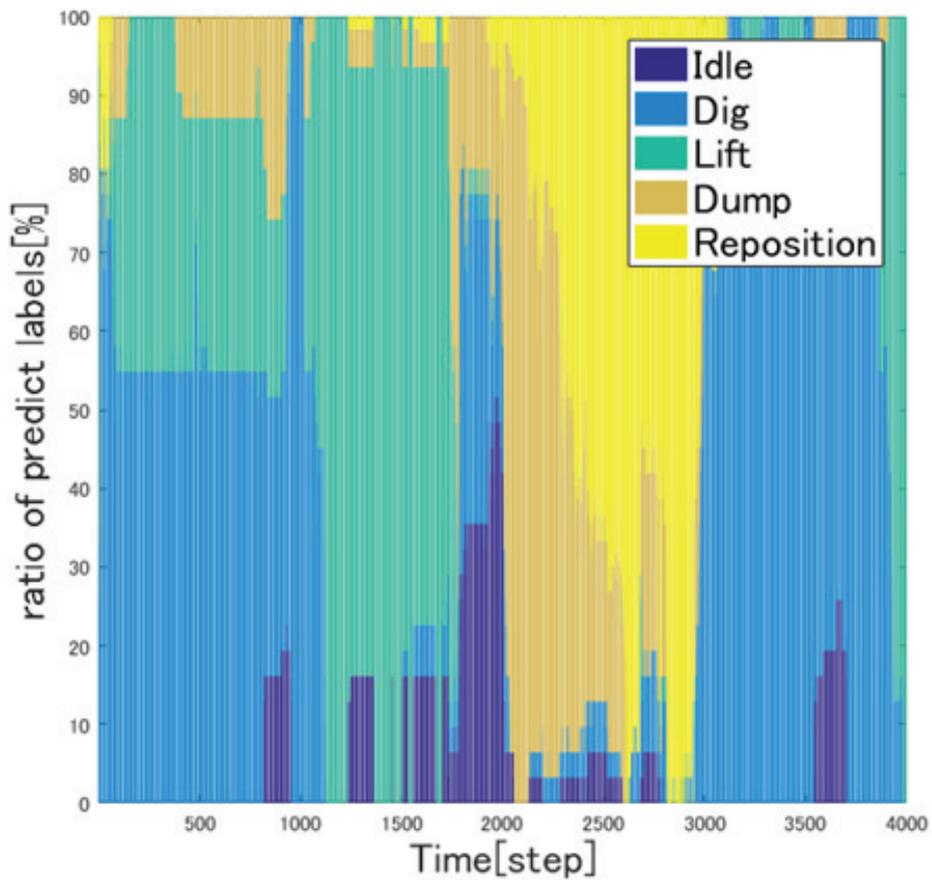


図 3.5: 一般作業者の分類結果

表 3.2: RF の有効性検証

	熟練者	一般	非熟練
平均	0.15	3.45	13.3
分散	2.92×10^{-5}	1.48×10^{-2}	2.06×10^{-2}
最大	0.26	6.74	15.9
最小	5.00×10^{-2}	1.70	10.5

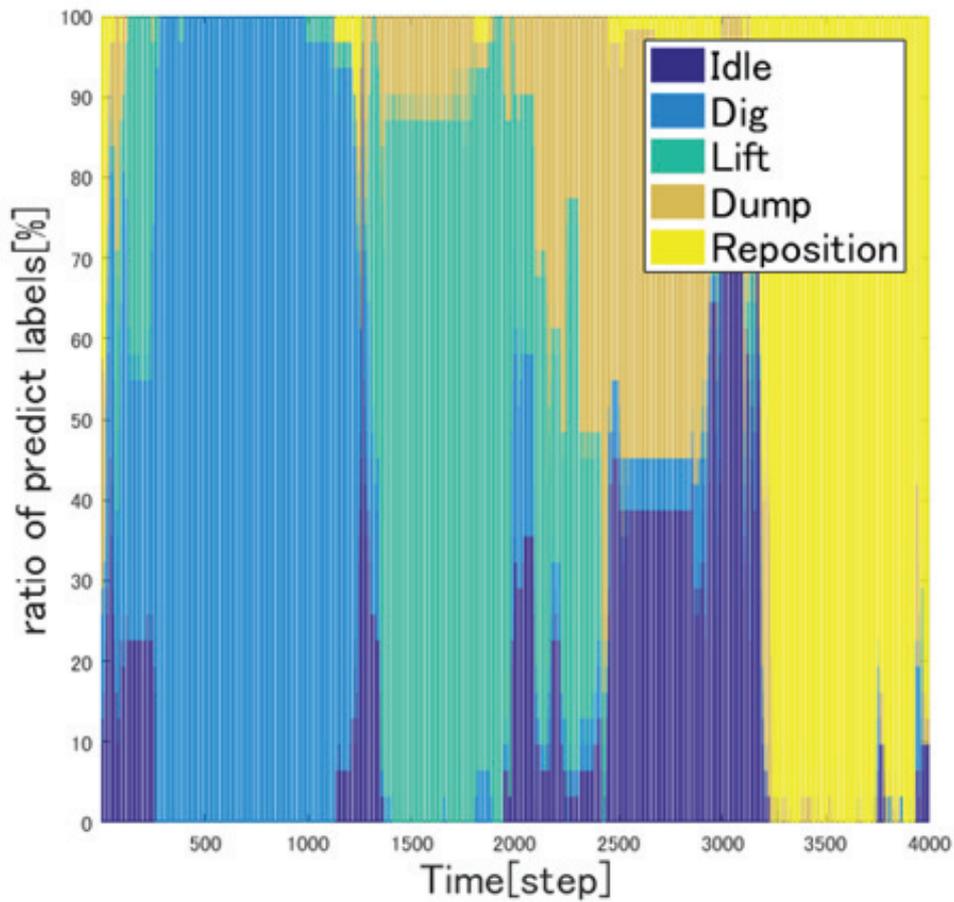


図 3.6: 非熟練者の分類結果

くなる。つまり、作業の状態を「待機」状態と判定した決定木の割合は、その作業における操作の不足量を示しており、作業の効率化、時間短縮を考えるうえで重要な操作スキルを定量化した指標として扱うことができる。

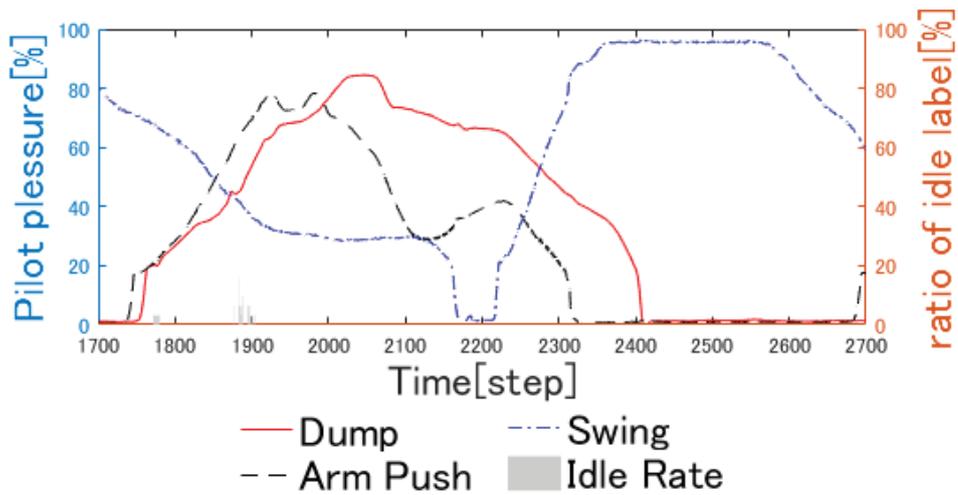


図 3.7: 熟練者の抽出結果

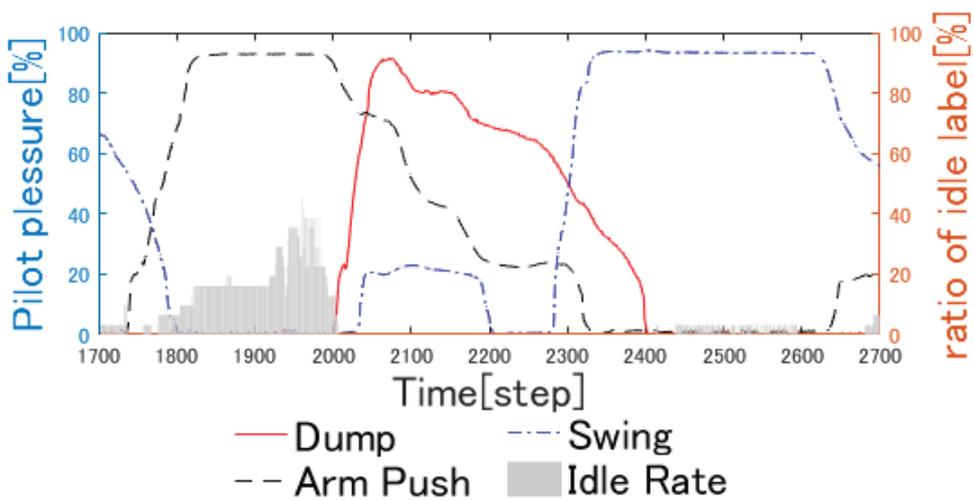


図 3.8: 一般作業者の抽出結果

3.4 状態遷移モデル表現

上記で述べた RF を使用して、基準となる操作パターンを抽出し、操作入力を特徴ベクトル、作業状態をラベルとして RF を学習する。学習

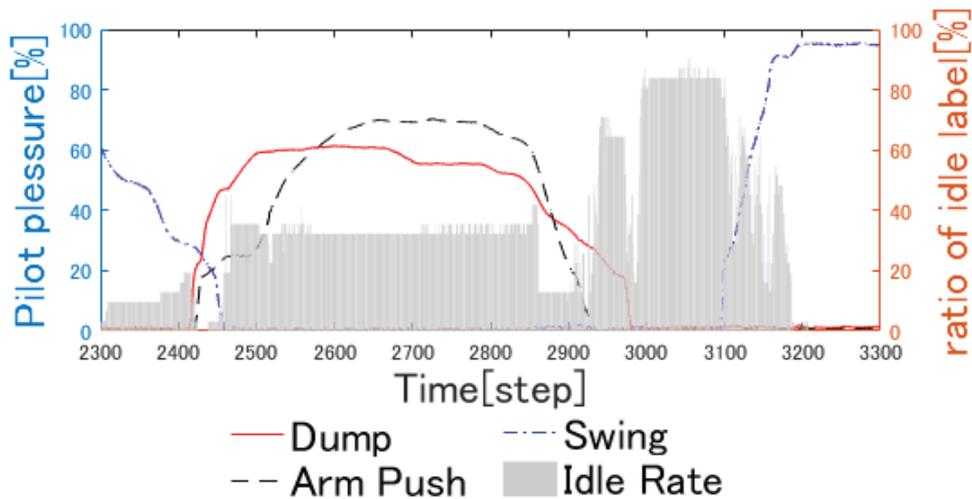


図 3.9: 非熟練者の抽出結果

された RF に対して、別の作業者の特徴ベクトルを入力し、出力された予測ラベルの割合を計算する。その出力された予測ラベルの割合を比較することで操作の差異を考察する。さらに、出力された予測ラベルの割合を状態遷移モデルで操作の差異を表現する。

3.5 状態遷移モデルの検証実験

計測するパイロット圧は、積込掘削操作に必要な「バケット掘削」、「バケット排土」、「アーム押し」、「アーム引き」、「ブーム上げ」、「ブーム下げ」、「旋回右」、「旋回左」の 8 つの操作を計測する。以降、計測されたパイロット圧はすべて、最大値を 100% として正規化された値を表示する。

3.5.1 実験手順

一般的な油圧ショベルの積込掘削時の作業状態は、「掘削 (Digging)」「持上旋回 (Lifting)」「排土 (Dumping)」「復帰旋回 (Repositioning)」の4つの状態に分類される。本研究の操作対象は、定量化に示す模擬積込掘削作業とする。これら4つの状態の一連の動作を1回行うことを、模擬積込掘削作業の1サイクルとして定義する。さらに、これら4つ状態に加えて、操作をしない状態を「待機」状態として、5つの状態に作業を分類する。実験手順は以下の通りである。

STEP 1 データ取得

非熟練者、一般作業員、熟練者のそれぞれの土砂のダンプ積み込み作業のパイロット圧データを取得する。

STEP 2 RF 学習

熟練者のパイロット圧データを特徴ベクトルとして、作業状態を表すラベルでRFを学習する。

STEP 3 定量化

得られた学習済みのRFを用いて、非熟練者、一般作業員、熟練者の操作入力に対して全サイクルで分類し、予測ラベルの割合を出力する。予測ラベルの割合から操作の特徴を分析する。

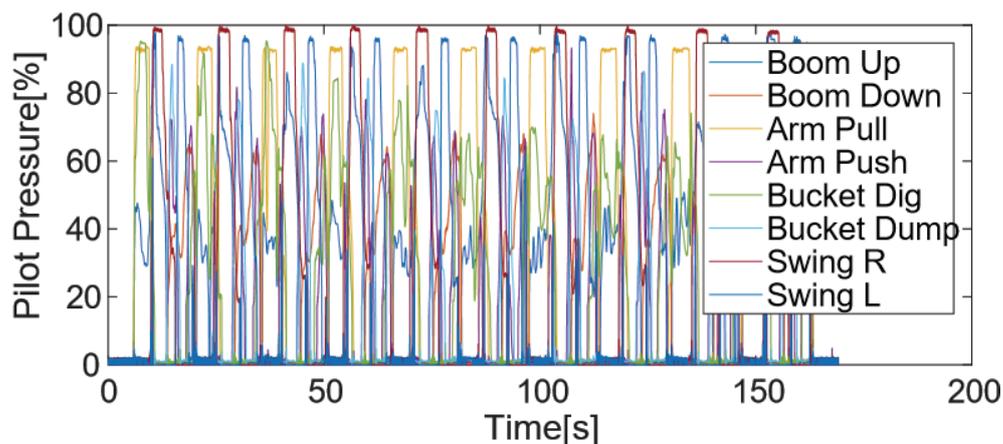


図 3.10: 熟練者の分類結果

3.6 検証結果

3.6.1 取得データ

基準とする熟練者は1名、非熟練者と一般作業者は各2名に対して、図3.3の作業を10サイクル実施し、各パイロット圧のデータを取得した。なお、非熟練者と一般作業者のそれぞれの結果は、同様の結果が得られていることから、以後のデータは各1名ずつのデータを示す。取得した結果を図3.10-3.12に示す。ここで、非熟練者は、建設機械の資格を有し、操作することは可能であるが、通常業務で使用していない者、一般作業者は、日常的に建設機械で作業を行っているが、機体操作の指導などできない者、熟練作業者は、勤務日のほぼ毎日で建設機械を操縦し、機体操作の指導が可能な者をそれぞれ示す。

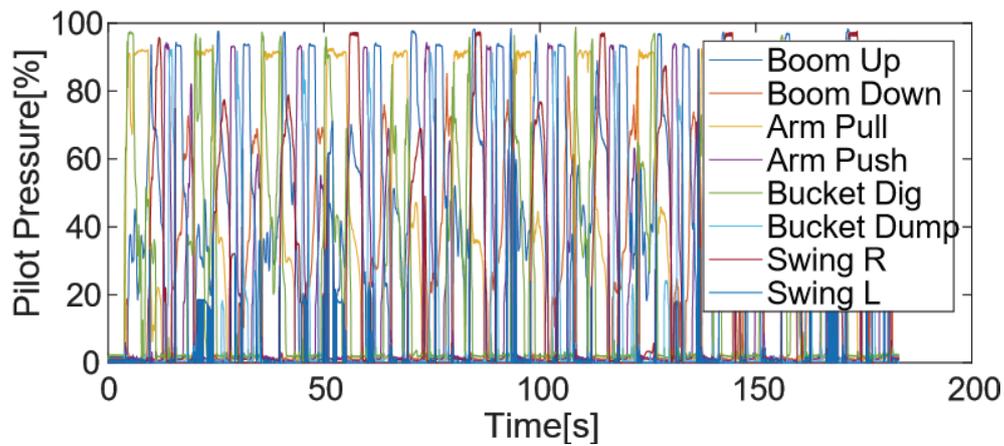


図 3.11: 一般作業者の分類結果

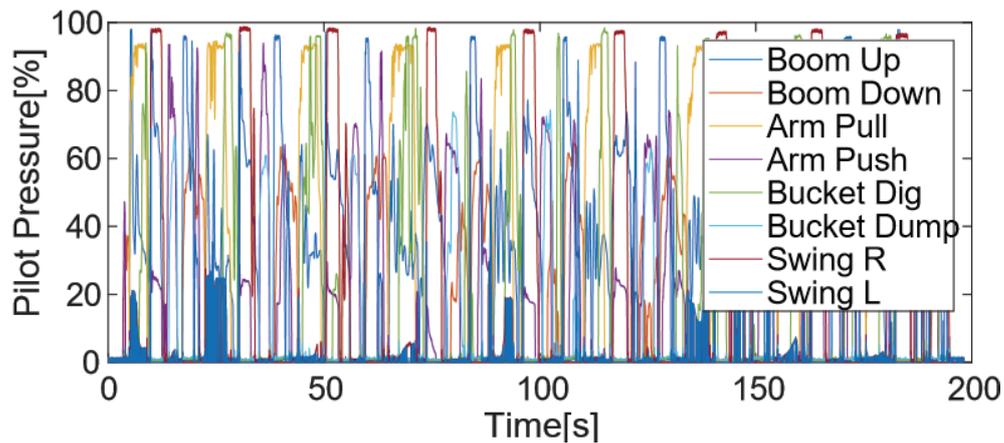


図 3.12: 非熟練者の分類結果

3.6.2 ランダムフォレストの学習

RF の学習は、熟練作業者の 1 サイクル分 (約 15 秒程度) のデータを用いる。熟練作業者の 1 サイクルのデータに対して、「待機」、「掘削」、「持上旋回」、「排土」、「復帰旋回」のラベリングを手動で行い、それらを訓練データとして RF に学習させた。学習データを図 3.13 に示す。学習時

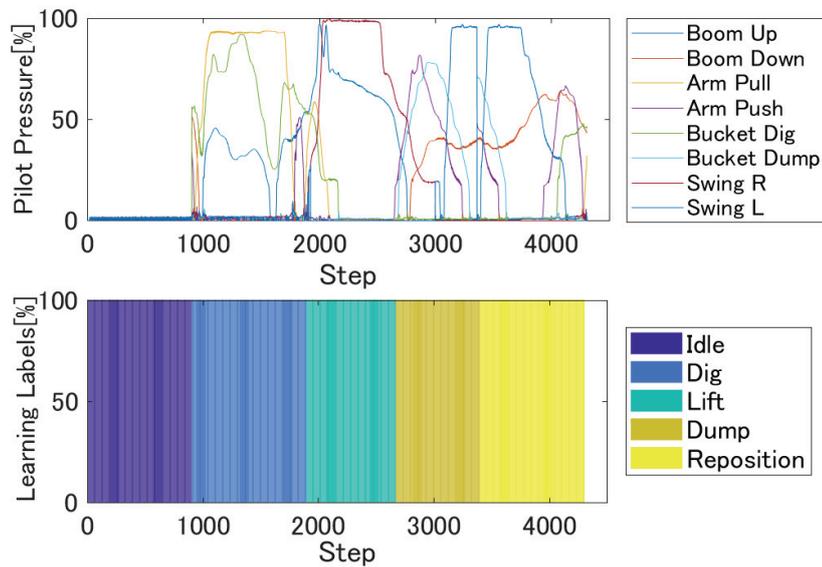


図 3.13: ランダムフォレストの学習データ

表 3.3: RF のパラメータ

決定木の数	学習データ数	決定木のパラメータ	
		学習データの抽出率	データの特徴量
31	4309	1/3	3

のパラメータを表 3.3 に示す。さらに、学習データをランダムに 10 分割にして、分割検証を実施することにより精度検証を行う。判定精度の結果を表 3.4 に示す。

熟練作業者の操作パターンを基準として、各作業者の作業を識別する。すなわち、上記学習結果を使用して、熟練作業者の特徴ベクトル、一般作業者の特徴ベクトル、非熟練作業者の特徴ベクトルに対して、作業状態をそれぞれ分類した。分類結果を図 3.14-3.16 に示す。ここで縦軸は、RF から出力されたラベルの割合を示し、横軸が時間軸を示す。

表 3.4: RF の有効性検証

予測ラベル	実際のラベル				
	待機	掘削	持上	排土	復帰
待機	99.8	0.1	0	0.2	0
掘削	0	98.7	0.1	0.1	0.6
持上	0	0.5	98.8	0.3	0
排土	0.1	0.1	0.2	75.4	19.5
復帰	0	0.7	0	24.1	79.8

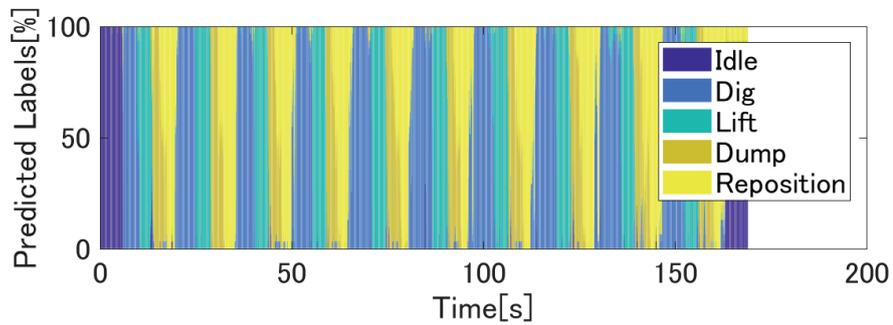


図 3.14: 熟練者の分類結果

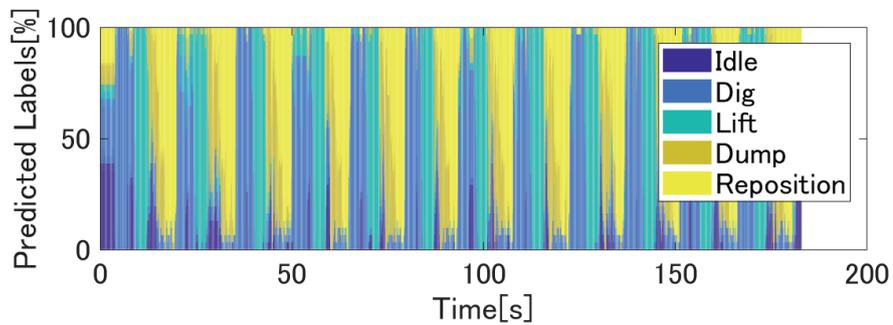


図 3.15: 一般作業者の分類結果

3.7 操作の特徴分析

分類結果の図 3.14-3.16 に基づいて、熟練者、一般作業者、非熟練者の差異を比較し、操作の特徴を分析する。

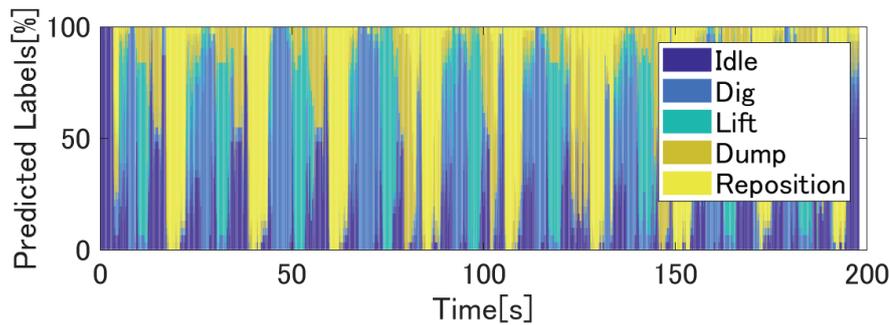


図 3.16: 非熟練者の分類結果

表 3.5: ラベルごとの面積比 [%]

	熟練	一般	非熟練
待機	0.37	3.49	15.25
掘削	34.48	31.36	27.36
持上旋回	23.76	25.51	17.94
排土	17.25	16.29	16.02
復帰旋回	24.15	23.35	23.42

3.7.1 行動モデルに基づく特徴分析

図 3.14-3.16 を定量的に比較するため、RF から出力されたそれぞれのラベルについて、横軸をラベル、縦軸を時間で表した棒グラフについて、各ラベルの面積比（以下、ラベルの面積比）を算出した結果を表 3.5 に示す。なお、作業前と作業後の無操作区間のデータは省いて算出されている。

表 3.5 の結果から、「待機」の割合が非熟練者、一般作業員、熟練者の順に小さくなっている。データや分析にランダム性を持つ RF による出力の有意性を検証するため、同じ訓練データと検証データを用いた RF に

表 3.6: 待機状態における RF の有効性の検証 [%]

	熟練	一般	非熟練
平均	0.15	2.71	13.6
分散	5.22×10^{-5}	5.21×10^{-3}	1.93×10^{-2}
最大	0.39	4.46	17.9
最小	3.87×10^{-2}	1.29	11.2

よる学習と分類を 50 回繰り返す。繰り返し計算の結果から、待機ラベルの面積比の違いについて、有意性を示した結果を図 3.6 に示す。

さらに、「待機」、「掘削」、「持上旋回」、「排土」、「復帰旋回」、各ラベルの割合の変化について、熟練者、一般作業員、非熟練者のそれぞれを図 3.17-3.19 に示す。図 3.18 3.19 から「待機」の割合が高くなるポイントを抽出すると、一般作業員も非熟練作業員も作業状態を切り替える際に「待機」の割合が高くなる傾向にある。ここで、「待機」のラベルは、作業を行っていない状態を定義していることから、「待機」が高いほど操作の入力が熟練作業員と比較して少ないことを示しており、100%に近い状態では、ほとんど操作していないことを示している。例えば、非熟練者は作業切り替わり時に「待機」が 100% 近くなることがあるが、このポイントは次の状態に移る前に、操作を止めていることがわかる。

ここで、一般的に人間が何等かの機械を操作する際の行動モデルを考えると、「認知」、「判断」、「操作」の 3 段階に分けることができることはよく知られている。このような機械操作の人間の行動モデルは「スキルベース」「ルールベース」「知識ベース」の 3 つの過程に細分化され、熟練度の高い作業員ほど、時間を要しない「スキルベース」で操作を行い、

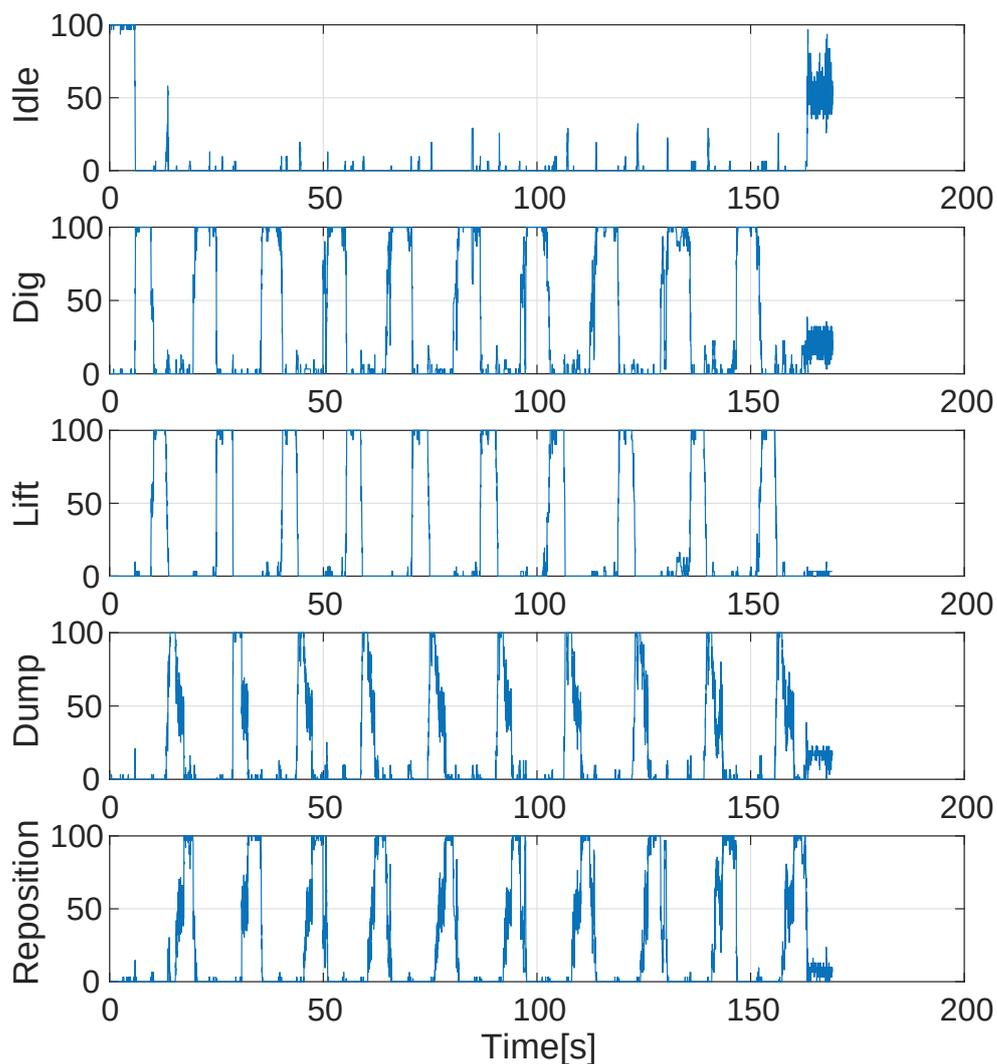


図 3.17: 熟練者のラベル割合の遷移

熟練度の低い作業者ほど、時間を要する「知識ベース」で機械操作を行っている [15]。この機械操作時の行動モデルに基づき、RF の出力結果を考察する。非熟練者は、「知識ベース」で操作を行うため、作業切り替え時には「判断」が必要となる。そのため、作業切り替え時には操作を緩める、または止めてしまうことで、作業に円滑さがなくなっているといえ

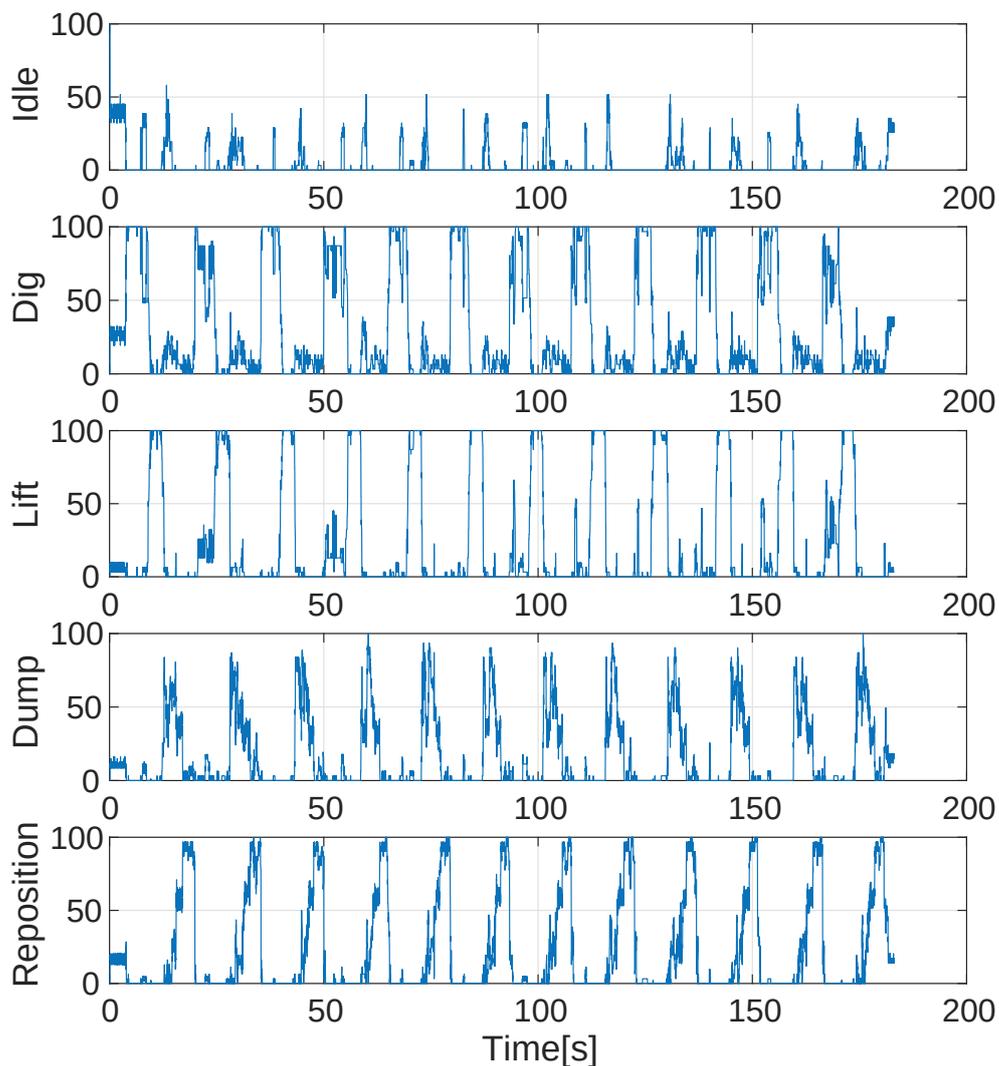


図 3.18: 一般作業者のラベル割合の遷移

る。一方、熟練者は、「スキルベース」操作として、「判断」を必要とせず、状態を「認知」することで、必要な操作量を入れ続け、円滑な作業を行っているといえる。また、一般作業者は、非熟練作業者のように「判断」するために操作を止めてしまうことはないが、熟練作業者と比較して、作業切り替え時に操作入力不足していると推察できる。

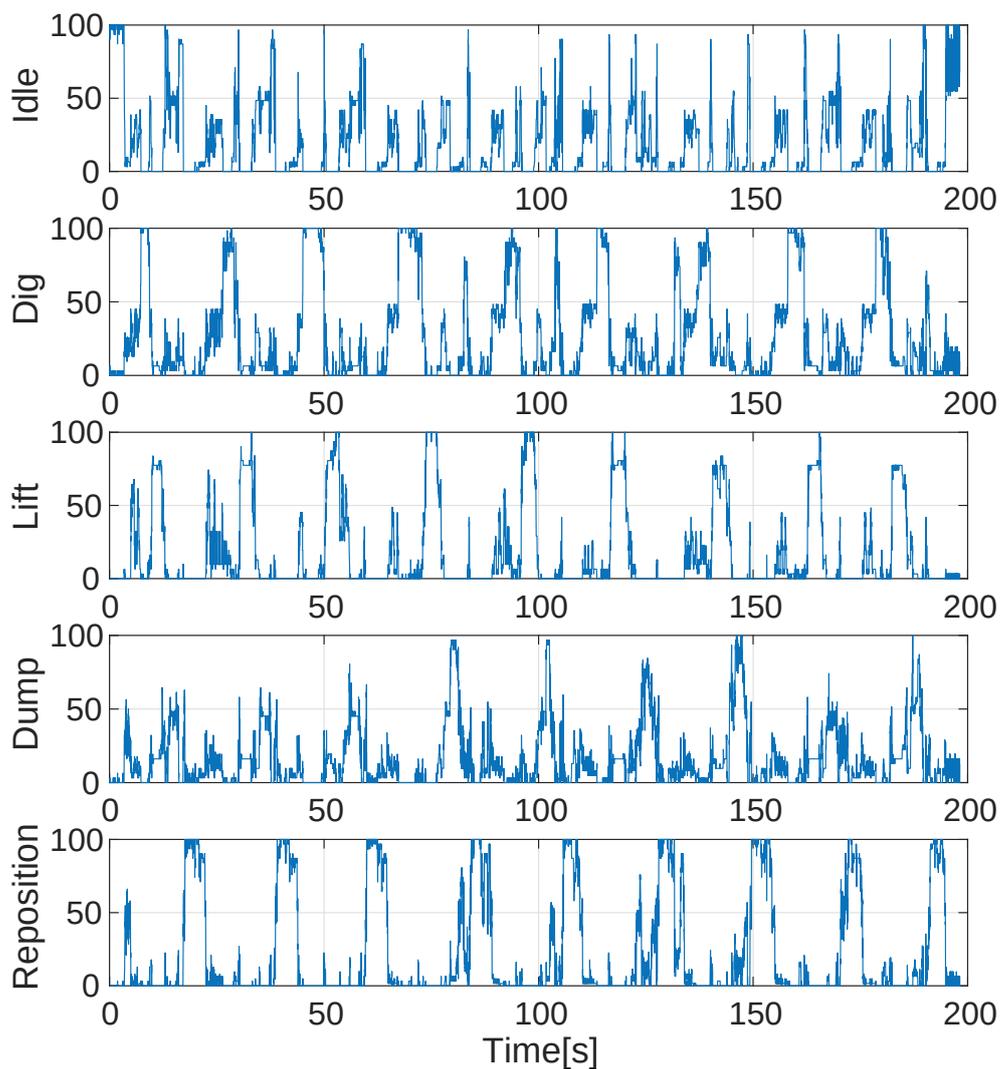


図 3.19: 非熟練者のラベル割合の遷移

3.7.2 状態遷移モデルに基づく特徴分析

RF から出力される 5 つのラベル「待機」「掘削」「持上旋回」「排土」「復帰旋回」に基づき、図 3.20 に示す状態遷移モデルを考える。この状態遷移モデルは、図 3.3 で示した模擬掘削の動作に基づき、「掘削」「持上

表 3.7: 状態遷移率 (熟練者)

遷移前の状態	遷移後の状態				
	待機	掘削	持上	排土	復帰
掘削	0.4	88.6	6.0	5.0	-
持上	0.4	-	95.6	3.8	0.2
排土	0.9	7.4	-	65.5	26.2
復帰	0.1	11.0	0.4	-	88.5

表 3.8: 状態遷移率 (一般作業者)

遷移前の状態	遷移後の状態				
	待機	掘削	持上	排土	復帰
掘削	4.8	68.1	9.3	17.0	-
持上	2.3	-	89.1	7.0	1.6
排土	8.2	13.2	-	47.3	31.3
復帰	3.3	12.3	1.2	-	83.2

巡回」「排土」「復帰巡回」を繰り返し、前の状態には遷移しないモデルと定義する。例えば、「掘削」から「持上巡回」には遷移するが、「掘削」から「復帰巡回」には遷移しない。また、「待機」は、いずれの状態からも遷移が可能とし、さらに「掘削」から「排土」への状態遷移といった、一つ前の状態以外への遷移も可能とする。なお、図 3.20 は図が煩雑になることを避けるため、次の状態への遷移と「待機」への遷移以外は矢印を記載していない。上記定義に基づき、RF から出力した各ステップの予測ラベルに対して、次のステップに対する状態遷移の割合について平均値を算出した。熟練者、一般作業者、非熟練者のそれぞれの結果を表 3.7-3.9 に示す。

表 3.7-3.9 に示した状態遷移の割合について、熟練者、一般作業者、非

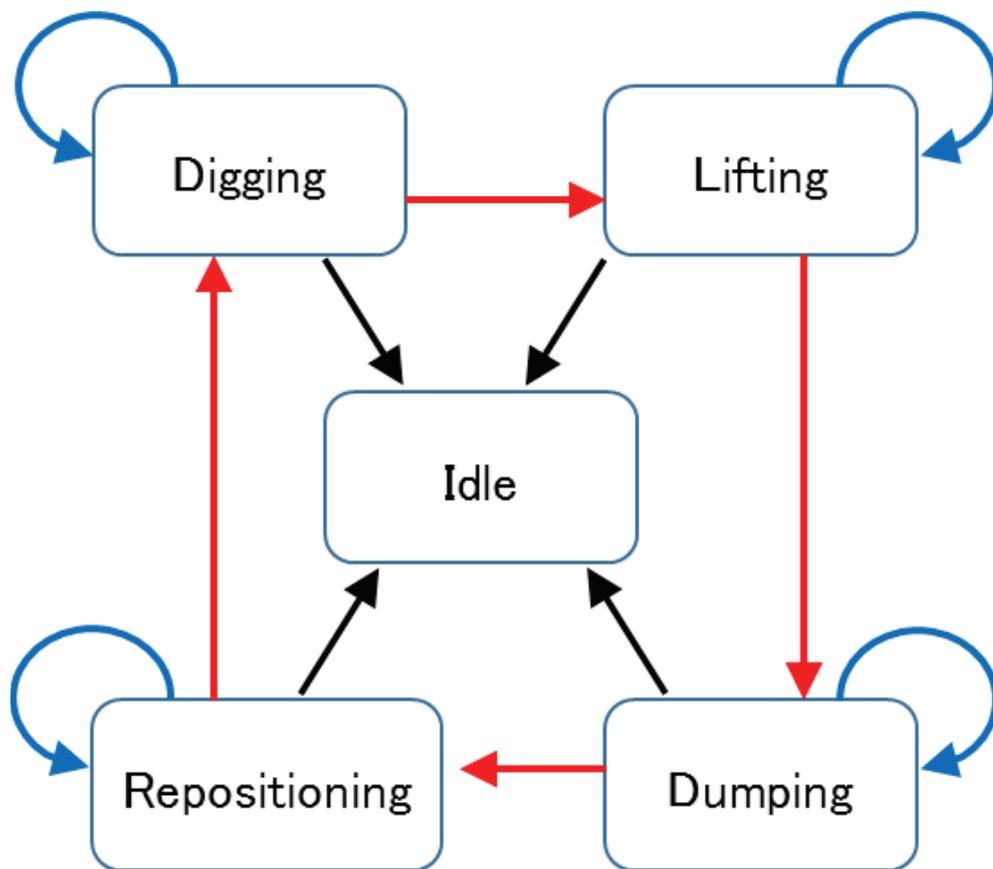


図 3.20: 模擬掘削の状態遷移

表 3.9: 状態遷移率 (非熟練者)

遷移前の状態	遷移後の状態				
	待機	掘削	持上	排土	復帰
掘削	20.4	49.5	15.2	14.9	-
持上	15.9	-	71.8	11.1	1.2
排土	18.3	18.3	-	45.1	18.3
復帰	6.3	11.7	1.3	-	80.7

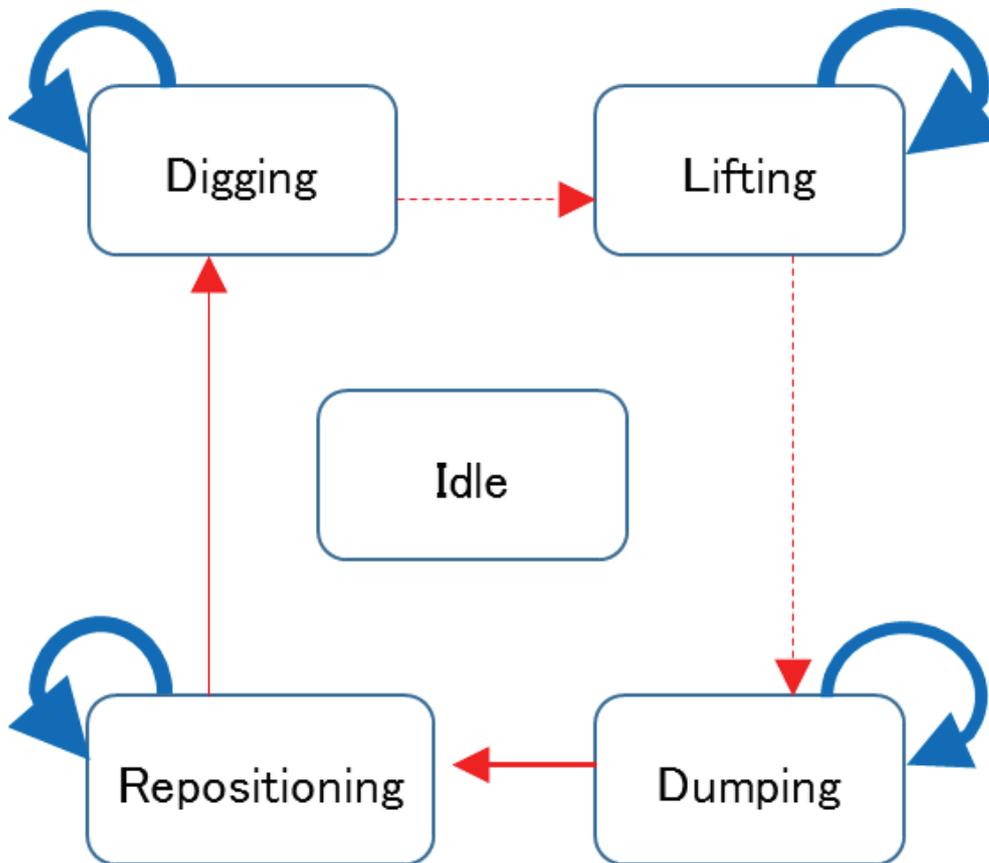


図 3.21: 熟練者の状態遷移

熟練者の差異を考察する。熟練者は、同一の状態への遷移割合が最も高く、次の状態への遷移割合がその次に高いことが分かる。また、待機への状態遷移がほとんどないことも明らかである。このように必要な4つの状態をスムーズに遷移させていることが分かる。次に、一般作業者は、待機割合は非常に低いものの、「掘削」から「排土」といった、次の状態以外への状態遷移割合が熟練作業者と比較して多いことが分かる。これは、基準として用いた熟練者の操作と比較して、RFで誤判定される要因の操作入力が含まれていることを示している。また、非熟練者は、前節

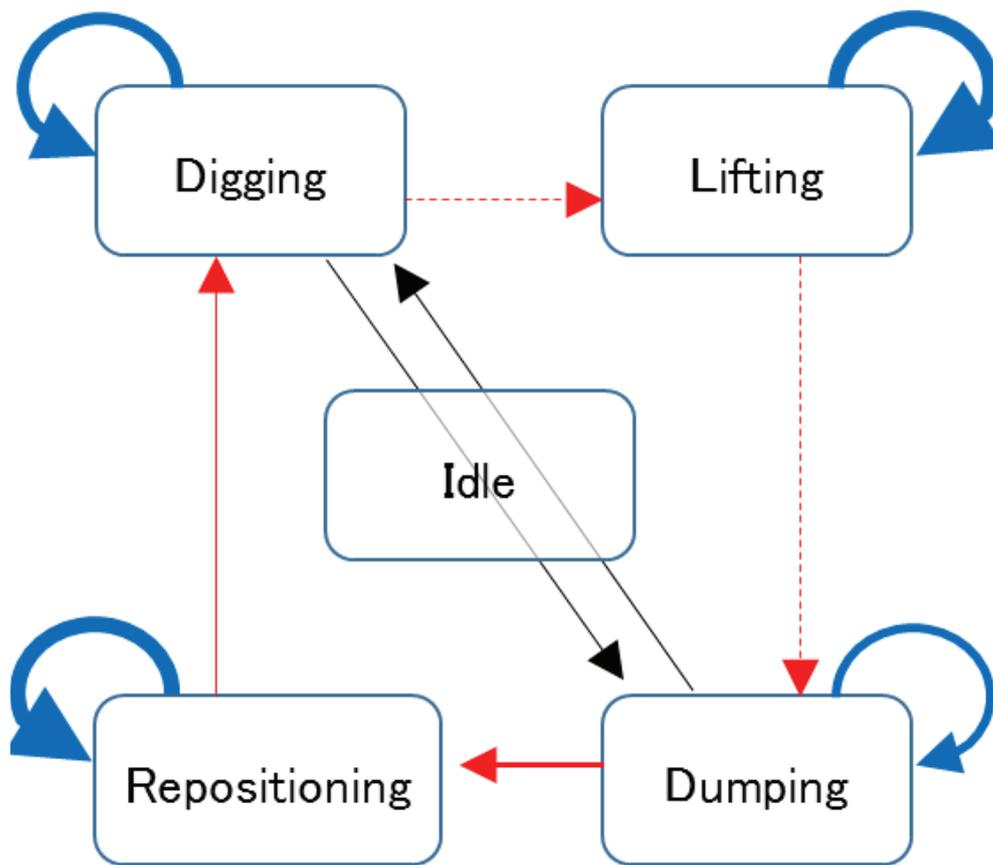


図 3.22: 一般作業者の状態遷移

で述べたことと同様に、「待機」への状態遷移が比較的多いことが明らかである。さらに、一般作業者と同様に、次の状態以外への遷移も比較的多く、熟練作業者の操作パターンと異なっていることを示している。これらの状態遷移を図示すると図3.21-3.23のように表される。なお、各図は表3.7-3.9に示される値の10%以下の数値は除外し、線の太さがおよその割合を示したものである。さらに、「掘削」「持上旋回」「排土」「復帰旋回」の遷移のうち、図3.20の矢印で10%以下の数値は点線で示す。このように、RFから出力された予測ラベルを使用して状態遷移モデル

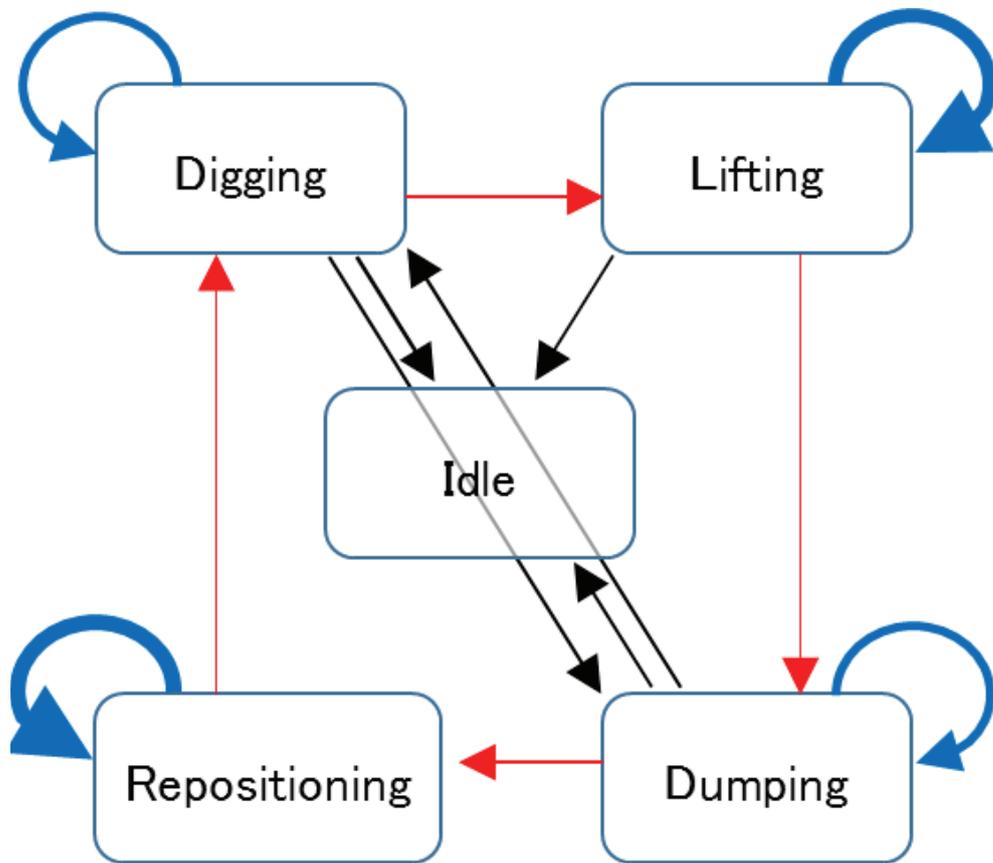


図 3.23: 非熟練者の状態遷移

で表現することにより，熟練作業者との差異を具体的に分析することができる。

3.8 結言

本章では，機械学習の一つである RF を用いた定量化手法を提案した。さらに，油圧ショベルの模擬掘削操作における特徴分析を通して手法の有効性を検証した。提案手法は，あらかじめ取得した熟練者の操作入力

を基準として、「掘削」「持上旋回」「排土」「復帰旋回」「待機」の5状態をRFに学習させる。その学習させたRFを使用して、熟練者、一般作業員、非熟練作業員の操作入力に基づいて作業状態を識別する。その識別された作業状態の差異により、操作の特徴を分析する。まず、RFから出力された予測ラベルの面積比に着目し、非熟練者の方が作業状態が「待機」と判定されることが多いことから、「待機」の判定タイミングを抽出することで、機械操作時の人間の行動モデルとの関係性について考察をおこなった。

次に、RFにより出力された予測ラベルに対して、次の状態に遷移する平均割合を算出し、状態遷移モデルを作成することで、操作特徴の差異を表現した。熟練者は、同一状態または次の状態への遷移の割合が非常に高く、一般作業員や非熟練作業員は熟練者と比較して、次の状態以外への遷移割合が多い。これは、基準とした熟練者の操作入力に対して一般作業員と非熟練作業員が異なる操作入力を行っているといえる。このような作業状態の差異を発生させる操作入力の差異を、熟練者の操作パターンに近づけていくことで、効率的な操作を行えるようになると思われる。

本手法は、システムが時間に応じて状態遷移する場合のモデリングと解析に有用な一方で、状態遷移によって表すことができない動的特性の考慮には課題がある。次章では、動的特性の考慮が可能な手法を、RFの重要度を利用して拡張することでこの課題を解決する。

4 ランダムフォレストによる変数選択に基づく データベース駆動型モデリングによる線形空燃 比センサ出力予測

4.1 緒言

本章では、動的特性に特徴のあるシステムの解析とモデリングについて述べる。3章では、システムの状態を複数定義し、システムの時間変化を状態遷移モデルとして表すことで、時間変化する特性をRFによって解析・モデリングする方法を提案した。本手法は、油圧ショベルの掘削作業のようにシステムが時間に応じて状態遷移する場合には有効だが、制御技術において重視されるシステムの動的特性の考慮には課題がある。そこで本章ではRFとシステムの動的特性を考慮可能な手法を連携させる手法を提案する。

1章で述べた通り、産業界のシステムの多くは複雑な非線形構造を有しており、モデリングを行うためには、この非線形性を考慮する必要がある。この問題に対して、ディープラーニングの一種であるRNNを用いた手法 [27] や非線形ブラックボックスモデリングによるシステム同定手

法 [28] など、様々な手法が提案されている。これらの手法は、大域的モデリングと呼ばれ、システムの入出力関係を単一のモデルにより表現する手法である。大域的モデリングは、システム全体の特性を、一つのモデルで表現するため、モデルの切り替え方法を別個に用意する必要がなく、モデルのチューニングも容易などの利点がある。一方で、モデルを構成するパラメータが難解であるため、モデルの構造から周波数解析に用いられる時定数やゲイン等のパラメータを得ることが一般的に難しく [29]、制御系設計や制御対象の解析等の応用は困難である。

このような問題に対して、局所モデルを逐次設計する手法である Just-in-Time (JIT) モデリング [29] [30] が提案されている。JIT モデリングは、蓄積したモデリング対象のデータをデータベースに蓄積し、そのデータベースを用いてモデリングする手法である。システムの局所的な特性を考慮したモデリングが可能であり、鉄鋼プロセスのモデリングなど、熱系プロセスに適用されている [31]。ほかにも、制御系設計分野では、制御器の逐次設計を行うデータベース駆動型モデリング (Database-Driven Modeling: DDM) 法が提案されており、局所モデルを利用した内部モデル制御系 (IMC) の設計手法 [34] や、一般化予測制御 (GPC) [33] に応用されている。一方で、DDM 法は一般的にシステムの入出力のようなシステムとの関係が明らかな入力変数を用いることが前提となるため、入力変数の中にシステムとの関連性が低い変数が多く含まれる場合は、それらの変数に過適合することで、モデリングの精度低下につながる場合がある。したがって、DDM 法のモデリング精度を向上するためには、システムに直接関連しない変数を特定・削除する必要がある。

入力変数とシステムの関連度合いを調べる手法として、統計的因果推論に基づく手法など、統計的アプローチを用いたが提案されている。しかしながら、これらの手法は設計パラメータを対象に合わせて個別に設計する手法を別途考えるなどの工夫が必要である。一方で、本手法で用いるRFの設計パラメータは、データ数や入力変数の数から求めることができる推奨値があり、比較的設計が容易である。また、2章で述べた通り、RFには変数の出力予測への貢献度合いを評価し、最も予測に貢献する変数を選択する機構があるため、入力変数の中にシステムとの関連性が低い変数がある場合でも、モデリングの精度維持が可能である。一方で、第3章で紹介したように、RFの適用先の多くは出力がラベル値となる分類問題であり、動的システムの解析やシステム同定など、出力が連続値をとるシステムへの適用は一般的に困難である。

そこで本論文では、DDM法にRFによる変数選択機構を付与したモデリング手法を提案する。RFに基づく変数選択は、相関などの線形な手法とは異なり、非線形性を考慮したうえで、変数の出力予測への貢献度合いの定量化が可能であり、変数の評価に有効である。本手法は、動的システムのモデリングをDDM法により行い、モデルに必要な入力変数の評価と選択をRFにより行うことで、システムに直接関連しない変数がある場合でも、モデリングの精度を維持可能な手法の確立を目指す。本章では、本手法の有効性を、まずは非線形システムのシミュレーションにより検証し、さらに後述する線形空燃比センサ(LAFS)の出力予測を本手法により実現することで、実システムにおける有効性を検証する。

4.2 ランダムフォレストの重要度に基づく変数選択法

4.2.1 変数選択法の概要

本章で対象とするシステムは図 4.1 に示すように、1つのシステム出力 $y(t)$ と、 D 個の入力変数 $x_1(t), x_2(t), \dots, x_D(t)$ をもつシステムとし、入力変数とシステムとの関係は未知とする。このとき、図 4.1 の対象システムは次式のようなになる。

$$y(t) = f(x_1, x_2, \dots, x_d) \quad (22)$$

ここで、 $f(\cdot)$ は非線形関数である。 x_1, x_2, \dots, x_d は、 $d(\leq D)$ 個のシステム出力 $y(t)$ に影響を与える変数を表しており、図 4.1 における変数セット (i) にあたる。また、図 4.1 の変数セット (ii) と (iii) は以下のとおりである。

(ii) 他の変数との相関が高い変数

$$x_{d+1}(t), x_{d+2}(t), \dots, x_{D'}(t)$$

(iii) システム出力 $y(t)$ との関連が低い変数

$$x_{D'+1}(t), x_{D'+2}(t), \dots, x_D(t)$$

ここで、 $D'(\leq D)$ は変数セット (i) と (ii) に該当する変数の総数である。図 4.1 に示すシステムのモデリング時、変数セット (ii) は、最小二乗法により過適合したシステムパラメータが算出される原因となり、変数セット (iii) は、DDM の近傍選択アルゴリズムにおいて、不適切な近傍を選択する原因になると考えられる。したがって、モデリングの精度を向上させるためには、あらかじめ変数セット (ii) と (iii) を特定し削除する必

要がある。まずは、変数セット (ii) を特定し削除することを考える。変数セット (ii) は他の変数との相関が高い変数であるため、すべての変数間の相関を調べ、相関が高い変数の対を探し出し、そのうちの1つを変数セット (ii) の変数として削除する。このとき、モデリング精度に貢献する変数の削除を避けるため、システム出力との相関係数がより低い変数を変数セット (ii) とする。

変数セット (iii) は、システム出力との関連性が低い変数であるため、RF の重要度により特定が可能である。具体的には、変数セット (iii) は重要度が低く算出されるため、低い重要度を持つ変数を、変数セット (iii) として削除する。

RF の重要度は、相関が高い変数が複数ある場合、本来の値よりも低く算出される場合がある。したがって、まず変数セット (ii) を削除し、その後変数セット (iii) を削除する。すなわち、以下の通りの手順となる。

[手順1] 変数セット (ii) の削除 (他の変数との相関が高い変数を削除)

[手順2] RF の重要度に基づく変数評価

[手順3] 変数セット (iii) の削除 (手順2において低く評価された変数の削除)

この手順により、最終的に変数セット (i) のみが残される。この変数セット (i) を変数選択の結果とする。

4.2.2 他の変数と高い相関をもつ変数の削除

前述の通り、DDM 法によるモデリングは、他の変数との相関が高い変数が存在すると、最小二乗法のアルゴリズムに多重共線性の問題が発

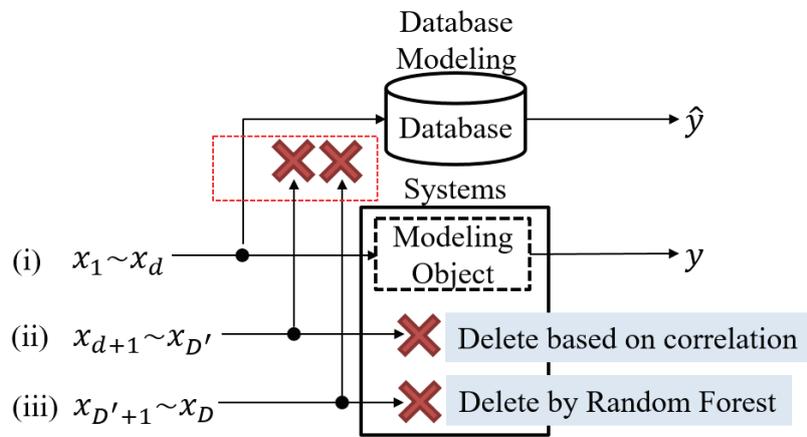


図 4.1: 対象システム概要図

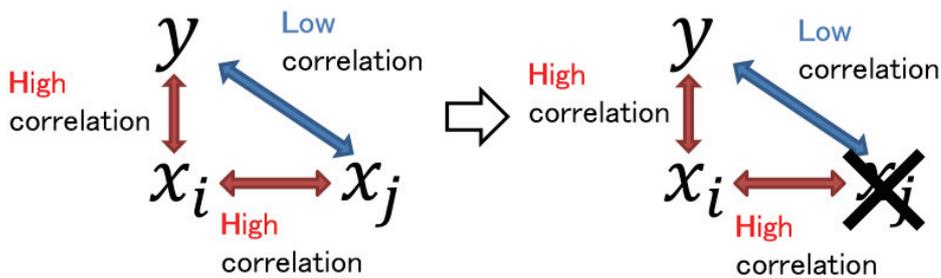


図 4.2: 他の変数との相関が高い変数の削除

生し、十分な精度のモデリングが難しくなる [37]。したがって、本手法でDDM法によってモデルを作成する前に、相関の高い一対の変数の組の一方を変数セット(ii)として削除する。

変数セット(ii)の削除手順をまとめた概要図を図4.2に示す。全ての入力変数の対の組み合わせから、図4.2における x_i, x_j に示すような相関が高い入力変数の対を見つけ、 x_j のようにシステム出力との相関が低い変

数を変数セット (ii) として削除し、もう一方の変数 x_i を変数セット (i) の候補として保存する。これにより、多重共線性の問題が解消し、モデリング誤差が低減される。

4.2.3 重要度に基づく非線形性を考慮した変数選択

先述の通り、システム出力との関連が低い変数セット (iii) は、DDM 法の近傍選択の精度を下げる原因となるため、除去する必要がある。したがって、本節では、RF により算出した変数の重要度を用いて変数セット (iii) を特定し削除する。具体的には、変数セット (i) と (iii) によって構成される変数から、DDM 法によるモデリング誤差が最も少なくなるように重要度上位 d 個の変数を選択し、上位 d 個を変数セット (i) として保存し、それ以外を変数セット (iii) として削除することで、DDM 法のモデリング誤差を最小化する。 d の計算過程で現れる d の暫定値を d' とする。このとき、 d' のとりうるすべての値について下記のような手順を実行する。

[Step1] 検証用データを用意し、すべての入力変数について重要度を算出する。

[Step2] 入力変数のうち、重要度上位 d' 個にあたるものを用いて、DDM 法を実行する。

[Step3] Step1 で得られたモデルを用いて、検証用データの出力予測を行い、予測誤差を誤差指標である平均二乗偏差

(Root Mean Squared Error:RMSE) を用いて算出する。

$$J_{\text{RMSE}} = \sqrt{\frac{1}{N} \sum_{k=0}^N (\hat{y}(k) - y(k))^2} \quad (23)$$

ここで、 J_{RMSE} は平均二乗偏差、 N は DDM 法のモデリング精度の検証用データ数、 $y(k)$ は検証用データに含まれているシステム出力の真値、 $\hat{y}(i)$ は予測出力である。以上の操作を d' の値を変えながら繰り返すことによって、 d' を探索する。平均二乗偏差が最小になる d' を d とし、変数セット (i) と変数セット (iii) を分割する。

4.3 データベース駆動型モデリング (DDM)

本節では、前節までの操作で選択された入力変数 x_1, \dots, x_d を基に、DDM 法によるモデリングを実行する方法について述べる。DDM 法は、動作点周りの特性を局所的に線形近似することで、システムを動作点ごとの線形モデルで表す手法である。DDM 法は、蓄積された過去の操業データを修正しながら学習を行うため、学習の際に対象のシステムの操業データが必要となる。したがって、対象システムから収集したデータを蓄積して、データベースを作成する必要がある。

4.3.1 対象システムの記述

モデリング対象のシステムが、以下に示す離散時間システムで表されるとする。

$$y(t) = f(\phi(t-1)) \quad (24)$$

ここで、 t はシステムの時刻、 $y(t)$ はシステム出力、 $f(\cdot)$ は非線形関数を表している。 $\phi(t-1)$ はシステムが時刻 t より以前にとった状態を表しており、以降本論文では情報ベクトルと呼ぶこととする。情報ベクトル $\phi(t-1)$ は次式で定義される。

$$\begin{aligned} \phi(t-1) = [& -y(t-1), \dots, -y(t-n_y) \\ & x_1(t-k_1-1), \dots, x_1(t-k_1-n_1), \\ & \dots, \\ & x_d(t-k_d-1), \dots, x_d(t-k_d-1-n_d)]^T \end{aligned} \quad (25)$$

ここで、 d は入力変数の次元数、 $x_i(t) (i=1, 2, \dots, d)$ は時刻 t における入力変数であり、4.2.1節における変数セット(i)となる。 k_i は入力変数 x_i のむだ時間である。 n_i, n_y はシステムの入出力 x_i, y の次数である。いま、式(24)で表される非線形システムが、局所的に次式のような線形モ

デルで表すことができると仮定する。

$$A(z^{-1})y(t) = \sum_{i=1}^d z^{-(k_i^{min}+1)} B_i(z^{-1})x_i(t) + \frac{\xi(t)}{\Delta} \quad (26)$$

ここで、 $\xi(t)$ はモデリング時の残差を表しており平均0、分散 σ^2 のガウス分布に従っているものとする。さらに、 z^{-1} は $z^{-1}x(t) = x(t-1)$ のように、時間遅れ作用素を意味しており、 $\Delta(=1-z^{-1})$ は差分演算子を表している。 k_i^{min} は最小のむだ時間推定値をあらわしている。また、システムパラメータ $A(z^{-1})$ と $B_i(z^{-1})$ は次式のとおりである。

$$A(z^{-1}) = 1 + a_1 z^{-1} + \dots + a_{n_y} z^{-n_y} \quad (27)$$

$$B_i(z^{-1}) = b_{0i} + b_{1i} z^{-1} + \dots + b_{n_i} z^{-n_i} \quad (28)$$

むだ時間が不明な場合は $k_i^{min} = 0$ とし、かわりに (28) 式の次数 n_i を通常よりも大きく設定することにより、むだ時間の不確かさを $B_i(z^{-1})$ に吸収させることでむだ時間の推定誤差の影響を最小限にとどめる。

4.3.2 システムパラメータの同定

パラメータ同定には式 (25) で表される情報ベクトル $\phi(t-1)$ を用いる。情報ベクトルはデータベースに保存し、必要に応じてパラメータ同定に利用する。情報ベクトルのなかでも、現在時刻を t としたとき、 $y(t+1)$ の算出に必要な情報ベクトル $\phi(t)$ は要求点と呼ばれる。このとき、式 (26) で示されたモデルは情報ベクトルを用いて以下のように表現することができる。

$$y(t) = \boldsymbol{\theta}^T(t-1)\boldsymbol{\psi}(t-1) \quad (29)$$

$$\boldsymbol{\theta}(t-1) := [a_1(t-1), \dots, a_{n_y}(t-1), \\ b_{01}(t-1), \dots, b_{n_d}]^T \quad (30)$$

$$\boldsymbol{\psi}(t-1) := [-y(t-1), \dots, -y(t-n_y) \\ x_1(t-k_1^{\min}-1), x_d(t-k_d^{\min}-1-n_d)]^T \quad (31)$$

このとき、(29) 式から、パラメータベクトル $\boldsymbol{\theta}(t-1)$ は次式で与えられる。

$$\boldsymbol{\theta}(t-1) = h(\bar{\boldsymbol{\phi}}(t-1)) \quad (32)$$

$$\bar{\boldsymbol{\phi}}(t-1) := [-y(t), -y(t-1), \dots, -y(t-n_y), \\ x_1(t-k_1^{\min}-1), \dots, x_d(t-k_d^{\min}-1-n_d)]^T \quad (33)$$

ただし、 $h(\cdot)$ は $\bar{\boldsymbol{\phi}}(t-1)$ をパラメータベクトル $\boldsymbol{\theta}(t-1)$ に変換する線形関数である。これは、 $\bar{\boldsymbol{\phi}}(t-1)$ を基に、 $\boldsymbol{\theta}(t-1)$ を算出可能であることを表している。 $\bar{\boldsymbol{\phi}}(t-1)$ および $\boldsymbol{\theta}(t-1)$ は、後述する式 (34) のデータベースの構成要素となる。

4.3.3 データベース構築

DDM 法は、データベース内に格納されているデータと紐づけられているパラメータベクトルを用いてモデリングする。したがって、モデリ

ングを行うために、システム出力と入力変数、およびそれらに紐づけられているシステムパラメータを格納するデータベースを構築する必要がある。したがって本節では、DDM法におけるデータベースの構築方法について述べる。手順は以下の通りである。

[STEP 1] 初期データベースの作成

[STEP 2] 距離の計算と近傍の選択

[STEP 3] 局所モデルの作成

[STEP 4] データ修正

次節以降で、各ステップの詳細を述べる。

[STEP 1] 初期データベースの作成

先述の通り、DDM法はシステム同定の際にデータベースに格納されている情報ベクトルおよびそれに紐づけられているシステムパラメータを用いる。したがって、まずは式(29)から、モデリング対象のデータに対して、最小二乗法を用いてシステムパラメータ $\theta(j)$ を推定し、その結果を情報ベクトル $\bar{\phi}(j)$ とともにデータベース内に保存し、これを初期データベースとする。このとき、初期データベース $\Phi(j)$ 以下のように表される。

$$\Phi(j) = [\bar{\phi}(j), \theta(j)], j = 1, 2, \dots, N_0 \quad (34)$$

ここで、情報ベクトル $\bar{\phi}(j)$ は(33)式によりシステム出力と入力変数から得られ、パラメータベクトル $\theta(j)$ は式(30)を用いて情報ベクトルより得られる。 N_0 は初期データベースに蓄えられている情報ベクトルの数を表している。以上のように、データベースに $\bar{\phi}(j)$ と $\theta(j)$ を格納し紐

づけることによって、後述する [STEP 2] で紹介するように情報ベクトル $\bar{\phi}(j)$ を用いて、対応するパラメータベクトル $\theta(j)$ を探索することが可能となる。なお、初期データベースでは、最小二乗法により対象システムを線形システムに近似しているため、この段階ではまだ非線形性への対応は行われておらず、パラメータベクトルは $\theta(1) = \theta(2) = \dots, \theta(N_0)$ となっている。

[STEP 2] 距離の計算，近傍の選択

DDMでは、システムの過去データである情報ベクトル $\bar{\phi}(j)(j \neq t)$ から、システムの現在の状態を表す要求点 $\bar{\phi}(t)$ の近傍データを選択し、それらに紐づけられているパラメータベクトルを用いてモデルを構成する。したがって、[STEP 2] では、以下のように定義した要求点 $\bar{\phi}(t)$ とデータベース内の情報ベクトル $\bar{\phi}(j)$ との距離を用いて近傍を選択する。

$$d(\bar{\phi}(t), \bar{\phi}(j)) = \sum_l \left| \frac{\bar{\phi}_l(t) - \bar{\phi}_l(j)}{\max_m \bar{\phi}_l(m) - \min_m \bar{\phi}_l(m)} \right| \quad (j = 1, 2, \dots, N_{db}) \quad (35)$$

ここで、 $\bar{\phi}_l(j)$ はデータベース内の第 j 番目に位置する情報ベクトルの第 l 番目の要素を表している。 $\bar{\phi}_l(t)$ についても同様に、要求点の第 l 番目の要素を表している。また、 N_{db} は、要求点が得られた時点でデータベースに蓄えられている情報ベクトルの数を表しており、 $N_{db} = t - 1$ である。さらに $\max_m \bar{\phi}_l(m)$ は、データベースに格納されているすべての情報ベクトル ($\bar{\phi}_l(j), j = 1, 2, \dots, N_{db}$) の第 l 番目の要素の中で最も大きな値をもつものを表している。同様に $\min_m \bar{\phi}_l(m)$ は第 l 番目の要素の中で最

も小さな値をもつものを表している。データベース内に存在するすべての情報ベクトルに対して、式 (35) により要求点との距離 $d(\bar{\phi}(t), \bar{\phi}(j))$ を計算し、距離の近いものから N_e 個の情報ベクトルを要求点の近傍として選択する。この操作により、要求点と同じ局所的特性に従う情報ベクトルが選択される。

[STEP 3] 局所モデルの作成

次に、[STEP 2] で近傍として選択された N_e 個の情報ベクトルから、次式の重み付き局所線形平均法 (Linear Weighted Average:LWA) により要求点回りの局所的なシステムパラメータを導出する。

$$\boldsymbol{\theta}^{old}(t) = \frac{\sum_{i=1}^{N_e} w_i \boldsymbol{\theta}(i)}{\sum_{i=1}^{N_e} w_i} \quad (36)$$

ここで w_i は近傍中の第 i 番目の情報ベクトル $\bar{\phi}(i)$ を基にして導き出される重みであり、次式で与える。

$$w_i = e^{-d(\bar{\phi}(t), \bar{\phi}(i))} \quad (37)$$

さらに、式 (36) から算出されるパラメータベクトル $\boldsymbol{\theta}^{old}(t)$ を利用して、次式のとおり予測出力 $\hat{y}(t+1)$ が得られる。

$$\hat{y}(t+1) = \boldsymbol{\theta}^{oldT}(t) \boldsymbol{\psi}(t) \quad (38)$$

以上の手順から、要求点まわりの局所的特性に応じたシステムパラメータと、それを用いて算出した予測出力が得られる。

[STEP 4] データ修正

[STEP 4]では、システムパラメータをシステムの動作点ごとに調整することで、モデルの精度を向上させる、具体的には [STEP 3] で得られたパラメータベクトル $\theta^{old}(t)$ に対して予測誤差の大きさに応じた修正を行い、得られた新たなパラメータベクトル $\theta^{new}(t)$ をデータベースに格納する。このとき、修正に利用する誤差の評価規範は、以下のように定義される。

$$J(t+1) := \frac{1}{2} \epsilon^2(t+1) \quad (39)$$

$$\epsilon(t+1) := y(t+1) - \hat{y}(t+1) \quad (40)$$

パラメータの修正は、最急降下法により誤差の評価規範 $J(t+1)$ が減少する方向におこなう。パラメータ修正式はパラメータベクトルの偏微分を用いて以下のように表される。

$$\theta^{new}(t) = \theta^{old}(t) - \eta \frac{\partial J(t+1)}{\partial \theta(t)} \quad (41)$$

ただし、 η は学習係数を表しており、対象システムに合わせて若干の調整が必要である。

最後に、修正されたパラメータベクトル $\theta^{new}(t)$ と要求点 $\bar{\phi}(t)$ を次式で表されるベクトル $\Phi(t)$ としてデータベースに保存する。

$$\Phi(t) := [\bar{\phi}(t), \theta^{new}(t)] \quad (42)$$

以降は、上記の [STEP 2] から [STEP 4] を一定回数に達するまで繰り返す。これにより、誤差指標 $J(t+1)$ が最小となるようなパラメータベク

トルがデータベースに保存されるため、DDM法のモデリング精度が向上する。

4.3.4 システムパラメータ同定（オンライン同定）

次に、先ほど設計したデータベースを用いて、システムパラメータをオンラインで推定する。その方法は前節で説明した [STEP 2] から [STEP 3] と同様である。つまり、推定パラメータベクトル $\hat{\boldsymbol{\theta}}(t)$ と局所線形モデルの推定値 $\hat{y}(t+1)$ は次式で得られる。

$$\hat{\boldsymbol{\theta}}(t) = \frac{\sum_{i=1}^{N_e} w_i \boldsymbol{\theta}(i)}{\sum_{i=1}^{N_e} w_i} \quad (43)$$

$$\hat{\boldsymbol{\theta}}(t) := \left[\hat{a}_1(t), \dots, \hat{a}_{n_y}(t), \hat{b}_{01}(t), \dots, \hat{b}_{n_{ad}}(t) \right]^T \quad (44)$$

$$\hat{y}(t+1) = \boldsymbol{\psi}^T(t) \hat{\boldsymbol{\theta}}(t) \quad (45)$$

4.3.5 数値例

まず、シミュレーションにより本章で提案した手法の有効性を検証する。検証用システムは (46) 式で与えられるものとする。ここで、 y_1 および y_2 は Hammerstein モデルの出力であり、それぞれ (47) 式および (48) 式で表される。さらに、 $\alpha (0 \leq \alpha \leq 1)$ は、 y_1 と y_2 の混合割合を表している。 y_1, y_2 および α は観測不可であるとする。

$$y(t) = \alpha y_1(t) + (1 - \alpha) y_2(t) \quad (46)$$

$$\begin{aligned}
y_1(t) &= 0.6y(t-1) - 0.1y(t-2) + 1.2x_1(t-1) - 0.1x_1(t-2) \\
x_1(t) &= 1.5u_1(t) - 1.5u_1^2(t) + 0.5u_1^3(t)
\end{aligned} \tag{47}$$

$$\begin{aligned}
y_2(t) &= 0.6y(t-1) - 0.1y(t-2) + 1.2x_2(t-1) - 0.1x_2(t-2) \\
x_2(t) &= 1.0u_2(t) - 1.0u_2^2(t) + 1.0u_2^3(t)
\end{aligned} \tag{48}$$

検証用システムを含むシミュレーションモデルの概要を図 4.3 に示す。また、図中に示されている入力変数 $x_1 \sim x_7$ の詳細を表 4.3.5 に示す。ここで、Model 1 は (47) 式で表されるモデルであり、Model 2 は式 (48) で表されるモデルである。検証用システムは図 4.3 に示すシステムの一部であり、入力変数の中には、検証対象に影響を与えない変数も存在する。したがって、検証対象に影響を与えない変数を除去しつつ、モデリングを行う必要がある。RF の重要度による変数選択の有効性を検証するために、他の変数と相関の高い変数セット (ii) はあらかじめ削除されているものとする。表 4.3.5 の入力変数 x_3 は α の変更を示す変数であり、0,1 のいずれか 2 値をとる。 $x_3 = 1$ のとき $\alpha = 0.9$ とし、 $x_3 = 0$ のとき $\alpha = 0.3$ となる。すなわち、 α は直接観測できないものの、 x_3 に着目することで、システムの変動を検知することができる。以上のことをまとめると、本検証では、システム出力 y に影響を与える入力変数 x_1, x_2 および x_3 は変数セット (i)、その他の入力変数 $x_4 \sim x_7$ はシステム出力 y に関係しないため、変数セット (iii) に属することになる。変数セット (iii) は、DDM法

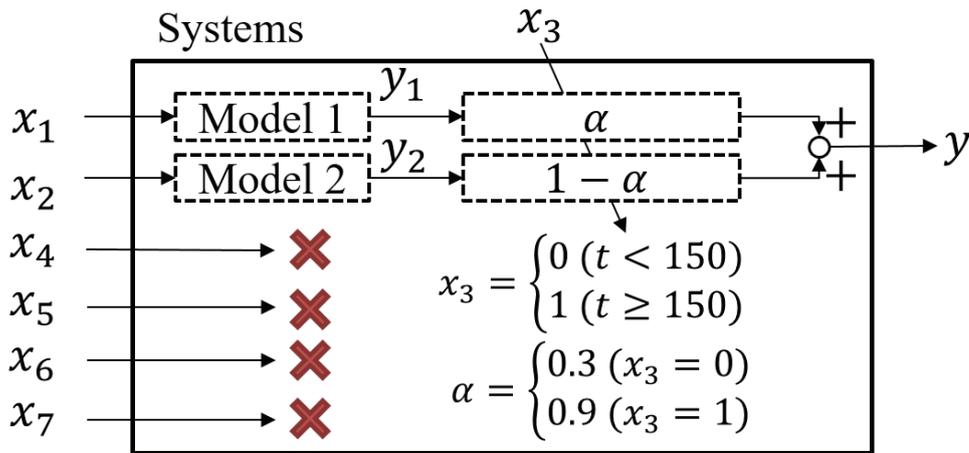


図 4.3: シミュレーションモデル概念図

表 4.1: シミュレーションに用いられる変数

変数	意味
x_1	Model 1 の入力変数 u_1 (変数セット (i))
x_2	Model 2 の入力変数 u_2 (変数セット (i))
x_3	α 切り替え信号 (変数セット (i))
x_4	分散 0.5, 平均 2 のガウス性白色雑音 (変数セット (iii))
x_5	無作為な矩形形状の信号 (変数セット (iii))
x_6	無作為な周期の sin 波 (変数セット (iii))
x_7	複数の無作為なインパルス信号 (変数セット (iii))

の近傍選択アルゴリズムに作用し、不適切な局所モデリングの原因となるため、DDM法によるモデリングの前に取り除く必要がある。変数セット (i) とシステム出力 y の関係は Hammerstein モデルを含む非線形であるため、本検証では変数のシステム出力への影響度合いを評価と、変数の評価を両立する必要がある。

本検証で用いる試験用学習データを図 4.4 に示す。 y はシステム出力、

表 4.2: パラメータ設定一覧

パラメータ名	設定値
決定木の数 N_T	30
条件分岐決定に利用できる変数の数	2
ブートストラップサンプリングの割合	1/3
勾配法によるパラメータ更新回数	100
勾配パラメータ η	0.01
近傍数 N_e	20
入力の次数 n_i	1
出力の次数 n_y	1

表 4.3: シミュレーションの予測誤差評価 (RMSE)

従来法 [33]	提案法
6.087	1.139

$x_1 \sim x_7$ は入力変数を示している。また、設計パラメータは表 4.3.5 の通りである。RF により算出した変数ごとの重要度を図 4.5 に示す。縦軸は重要度、横軸は変数の通し番号をそれぞれ示している。図中の赤枠で囲われている部分は、本手法により選択された変数である。図 4.5 の選択された変数をみると、変数セット (i) に属する入力変数 x_1, x_2 および x_3 が、すべて選択されており、かつ変数セット (iii) に属する入力変数 $x_4 \sim x_7$ はすべて除去されていることがわかる。したがって、提案法はシステムと入力変数の関係を適切に評価できているといえる。提案法による変数選択の有効性を検証するために、従来法としてシステム出力の相関が 0.6 以上である入力変数を選択しモデルを構築し、同様の検証を行った。なお、従来法では x_1 のみが選択された。従来法および提案法の適用結果を図 4.6 に示す。1 段目の赤色の破線は出力 y の真値である。また青

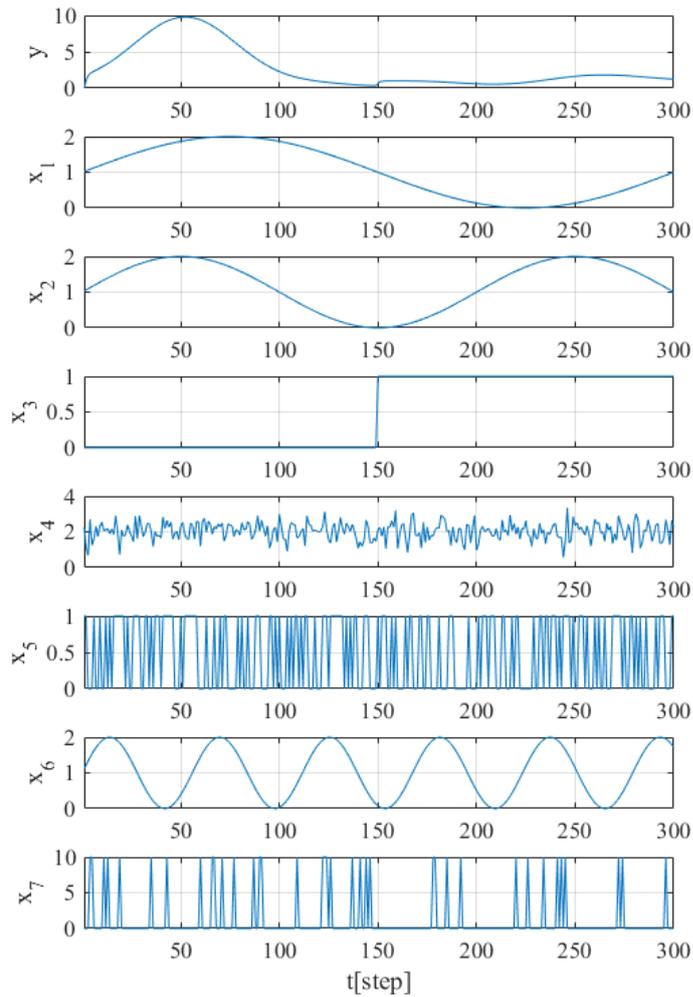


図 4.4: シミュレーションに用いる入力変数

色の実線が提案法，緑色の一点鎖線が従来法の予測した出力である。2 段目，3 段目はそれぞれ Model 1，Model 2 への入力，4 段目は予測誤差であり，出力と同様に青色の実線が提案法，緑色の一点鎖線が従来法の予

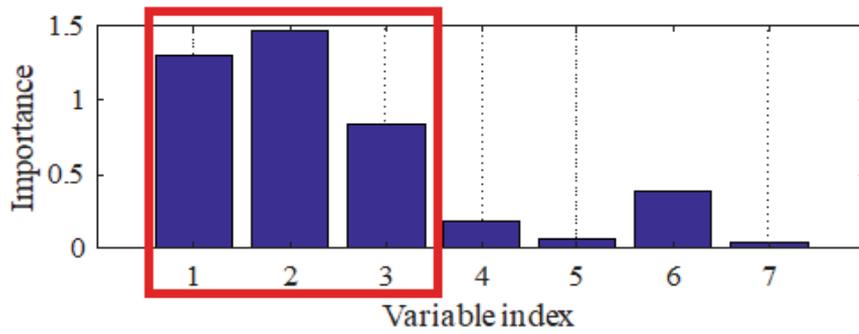


図 4.5: 入力変数の重要度

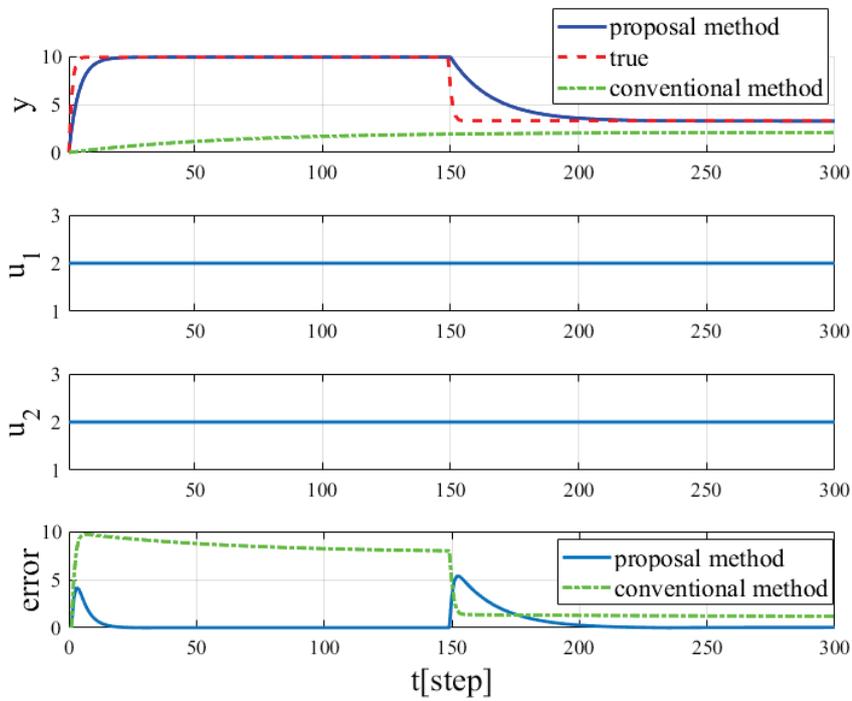


図 4.6: シミュレーション予測結果

測誤差である。さらに、従来法、および提案法のシステムパラメータの予測結果を図 4.7 に示す。また、図 4.6 の予測誤差を比較するために、式

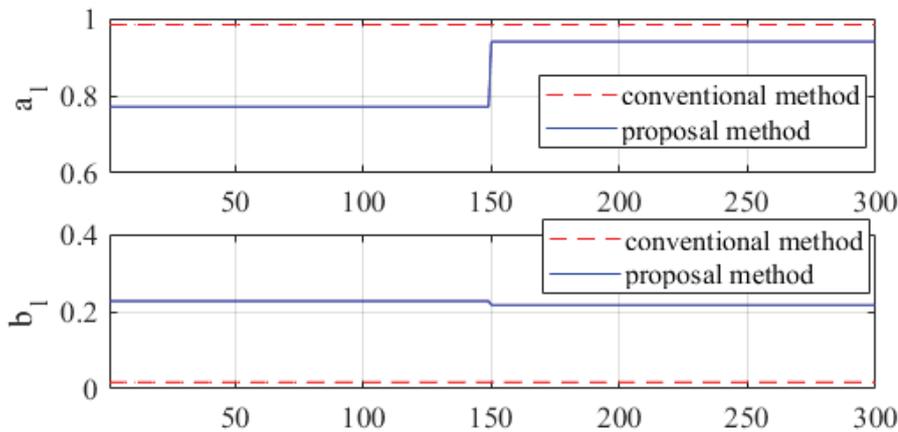


図 4.7: 推定システムパラメータの推移

(23) に示す誤差指標である RMSE の計算結果を表 4.3 に示す。図 4.6 と表 4.3 から、相関を用いて変数選択を行う従来法と比べて、RF の重要度により変数選択を行う提案法はより大きく予測誤差を低減していることがわかる。これは、図 4.7 のシステムパラメータの推移からわかる通り、従来法がシステム出力との相関の低い x_3 を選択していないことにより、 α の変動を十分に予測できていない一方で、提案法は x_3 をを選択することで α の変動を予測し、システムパラメータを更新したためである。以上のことから、本手法は、非線形システムに対して、状況に応じて相関の低い変数を考慮しながら、モデリング精度を向上させることができる。

4.4 線形空燃比センサ出力予測による検証

4.4.1 線形空燃比センサ出力予測

線形空燃比センサ (Linear Air Fuel Ratio Sensor:LAFS) は、車両に搭載されるセンサの一種であり、燃焼部に送り込まれる空気と燃料の混合比である空燃比 (A/F) をエンジンの排気から測るセンサである。車両とエンジン、および LAFS の関係の概要図を図 4.8 に示す。エンジンの燃焼時に排出される一酸化炭素 (CO) やノックス (NO_x) 等の有害物質を最小限に抑えるためには、エンジン内を最適な A/F に維持する必要がある。したがって、この LAFS により A/F を正確に観測しながら、エンジン制御を行う必要がある。しかしながら、LAFS は長期の走行により経年劣化し、図 4.8 のように A/F の変化に対する応答速度が大きく減少するため、安定した A/F の観測のためには、同じく車両にて取得可能な制御値を用いて、劣化した LAFS 出力を補正し、劣化前の値を推定する必要がある。ただし、LAFS 出力に関連する制御値が数個であるのに対し、車両から取得可能な制御値は数百種類あり、LAFS と関連する制御値の正確な種類と数は不明である。また、LAFS と他の制御値間の関係の多くは非線形であり、線形的な手法による LAFS 出力に関連する制御値の特定は困難である。したがって、LAFS 出力を予測するためには、非線形性の考慮と、制御値の取捨選択を両立する必要がある。

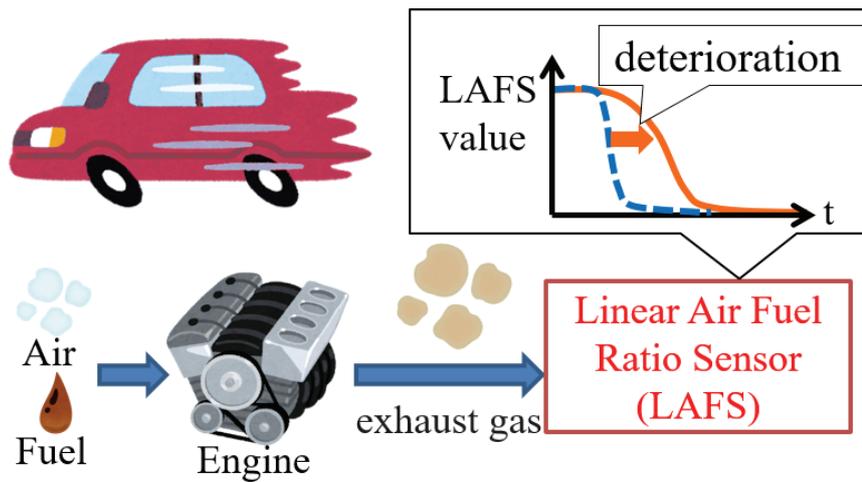


図 4.8: LAFS 概要図

4.4.2 検証結果

本検証で対象となった制御値（入力変数）は 347 種類であり、まずはこの中から変数セット (ii)(iii) にあたるものを除去する。相関 0.99 以上の変数の組について、4.2.1 節の [手順 1] を実行した結果、219 種類の変数セット (ii) の制御値が削除され、残りが 129 種類となった。ここからさらに変数セット (iii) にあたる制御値を RF に基づく変数選択により除去することで、変数セット (i) のみを残すことができる。[手順 2] により算出した制御値ごとの重要度を図 4.9 に示す。算出した重要度を基に [手順 3] を実行したところ、126 種類の制御値が変数セット (iii) として削除され、3 種類の制御値が変数セット (i) として残った。提案法の有効性を検証するために、システム出力と相関の高い変数を選択し、DDM 法によりモデルを構築する従来法と、RF の重要度に基づいて変数選択を行いながら DDM 法によりモデルを構築する提案法によりそれぞれ LAFS 出力

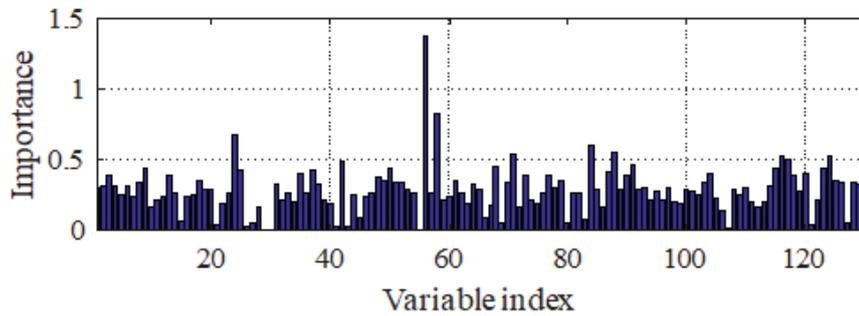


図 4.9: LAFS 予測における重要度

の予測を行った。予測結果を図 4.10 に示す。なお、従来法に用いる変数は、出力との相関が 0.6 以上と判定された 4 変数を選択した。上段のグラフは、縦軸の値を正規化した LAFS 出力値であり、赤色の破線が出力の真値、緑色の一点鎖線が従来法による予測値、青色の実線が提案法による予測値である。下段はそれぞれの予測誤差である。図 4.10 に示すように、従来法と比べて、提案法は予測誤差が小さいことがわかる。これは、従来法のデータ選択アルゴリズムが、相関が低い重要な変数を選択できていないためだと考えられる。一方、提案法では、相関の大小にかかわらず重要な変数を選択しているため、より高い予測精度が得られている。以上の結果から、本手法は、LAFS 出力予測のように、多くの入力変数が存在し、システム出力との関連が不明な場合でも有効であることがわかる。

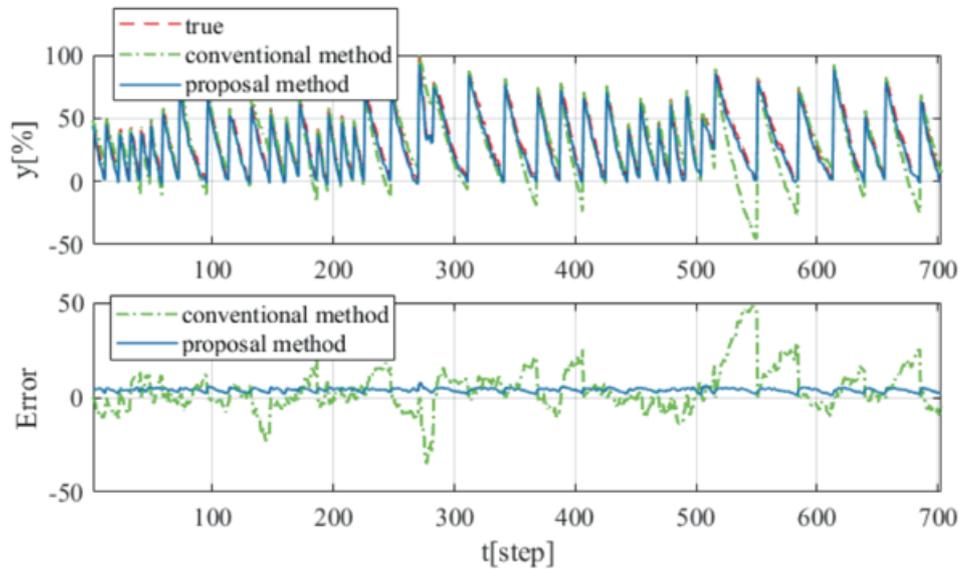


図 4.10: LAFS 出力予測結果

4.5 結言

本章では、RF の重要度に基づいた変数選択を行うことによって、DDM 法の改良を目指した。従来の DDM 法には、モデリングに利用する入力変数にシステムとの関連性の低い変数が多く含まれる場合に、適切な近傍選択を行えず、モデリング精度が低下する問題があった。

そこで本章では、RF の重要度を用いた非線形性へ対応した変数選択手法を提案した。RF は、一般的に連続値をとるシステムのモデリングに適さないが、モデル本体に DDM 法採用することで、この問題を解決した。また、本手法の有効性を、シミュレーションと実機データを用いた LAFS 出力予測への適用を通して行い、提案法は相関を用いて変数選択を行う従来法に比べ、LAFS 出力の予測誤差が低減されていることが分

かった。これは、LAFS 出力に関連しない変数の削除に加え、LAFS 出力との相関が低いものの、関連性の高い変数が、重要度により高く評価されたためだと考えられる。本検証により、3章で解説した状態遷移モデルを利用した解析に加えて、動的特性に特徴のあるシステムの解析とモデリングについても、RF を用いたモデリングとその応用によって実現可能であることが確認された。

5 結言

本論文では、時間変動する非線形システムのモデリングと、モデルの解析性の確保の両立を目指して、RFを軸にした2つの手法を提案し、シミュレーション、実機実験を投資手有効性を検証した。具体的には、下記の2手法について考察した。

- (1) システムの状態定量化・状態遷移モデルを用いた解析手法
- (2) 時系列データのモデリングと解析

まず、第2章では、前半で決定木の構造とアルゴリズムに触れたのち、RFについて述べた。RFは機械学習の一種であり、構造解析が比較的容易な決定木と呼ばれる学習機によって構成されており、本論文において重要な要素である非線形性の考慮と解析性の両立を担った。さらに、モデルの応用手法であるRFを用いた説明変数の重要度算出についても述べた。

第3章では、(1)について述べ。油圧ショベルの掘削作業の解析を通して、提案手法の有効性を確認した。本章で示した通り、解析に使用した操作入力の時系列データであり、それらを「掘削」「持上旋回」「排土」「復帰旋回」「待機」の5つの状態に分けて解析を行った。本章前半では、RF内に含まれる決定木を利用して、作業状態を定量化した。これにより、作業者の熟練度合いを操作入力の「待機状態」の割合として定量化

可能であることを確認した。本章後半では、「待機状態」以外の割合の時間的变化に着目し、状態遷移モデルに落とし込むことで、状態遷移の様子から「熟練者」「一般作業員」「非熟練者」の操作入力の差異をより詳細に評価した。これらの検証により、本手法を用いれば、一般的に時系列データの扱いが難しいとされる RF によって、時系列を考慮した解析が可能であることを確認した。

第4章では、(2)について述べた。動的特性に特徴のあるシステムについて、システムの動的特性と非線形性を考慮可能なデータベース駆動型モデリング (DDM) に、RF の解析機能を搭載することにより、解析とモデリングを同時に行うことで、モデルの精度を向上させた。また、提案手法の有効性を、シミュレーションと自動車の線形空燃比センサ (LAFS) の出力予測により検証した。結果として、347種類の入力変数から、3つの重要な変数を抽出し、DDM によるモデリングに利用することによって、モデリング誤差が大きく減少することを確認した。

以上により、提案手法のモデル化とその応用手法について有効性が確認された。

本論文では、RF の異なる特性を用いて、それぞれの課題を解決したが、これらの手法は最終的に統合可能であると考えられる。例えば、自動車の LAFS 出力について動的特性の解析とモデリングを行う際、その中の決定木に着目して状態遷移モデルに落とし込めば、その自動車の走行状態だけではなく、運転状況、道路状況など、自動車を取り巻く状態の違いを捉えることが可能となり、さらに実用的なモデリングと解析が実現できると考えられる。このように、状況に合わせたモデリングと解析は、

第1章で述べたスマートプラットフォームの構築につながり、最終的には高度な適用制御技術の基盤になりうると考えている。したがって、今後は提案手法の適用制御技術への応用を目指して研究を進めていきたい。

謝辞

まず、終始熱心にご指導を頂いた広島大学の山本透教授に心から感謝の意を表します。山本教授の研究への熱意と学生に向き合う姿から、研究・教育両面において大きな刺激を頂きました。また、多くの有益な助言とご指導を頂いた大阪公立大学の野修一教授、広島大学の林田智弘准教授、脇谷伸准教授に感謝いたします。

油圧ショベルの分析にあたり、操業データの提供と、実機操作のご協力を頂いたコベルコ建機株式会社の山崎洋一郎氏、上田浩司氏に感謝いたします。また、状態遷移モデルの作成について、熱心にご指導を頂いたコベルコ建機株式会社小岩井一茂氏に心から感謝いたします。

LAFS 試験データ取得にあたり、実機試験とデータ利用をご快諾頂き、共同研究の枠を超え多くの助言を頂いたマツダ株式会社の中川正氏、吉田昌弘氏、伊藤啓介氏には大変お世話になりました。ありがとうございました。

論文執筆にあたり、ご指導を頂いた広島大学の木下拓矢助教に感謝いたします。

広島大学の中本昌由助教のご助力に感謝いたします。また、システム制御論研究室の院生、学部生をはじめとした多くの方々に感謝いたします。

最後に、本論文の執筆にあたり、私を支えてくださったすべての方々に感謝いたします。ありがとうございました。

参考文献

- [1] Toru Yamamoto, Takuya Kinoshita: "Design of a Smart Adaptive Control System", International Symposium on Advanced Control of Chemical Processes, International Federation of Control, Whistler, Canada(2015)
- [2] Toru Yamamoto: "Design and Experimental Evaluation of a Data-Driven Smart Adaptive Control System", Computer Process Control Seminar in University of Alberta, Dept. of Chemical and Materials Engineering, University of Alberta, Edmonton, Canada(2015)
- [3] Geoffrey E. Hinton, Simon Osindero, Yee Whye Teh: "A fast learning algorithm for deep belief nets", Neural Computation Vol18, pp 1527 - 1554 (2006)
- [4] Leo Breiman: "Classification and Regression Trees", Taylor & Francis (1984)
- [5] YY Song, LU Ying: "Decision tree methods: applications for classification and prediction" Shanghai archives of psychiatry, Vol.27, No.2 pp.130-135 (2015)
- [6] Leo Breiman: "Random Forests" Machine Learning, Vol.45, No.1 pp.5-32 (2001)
- [7] Leo Breiman: "Bagging Predictors", Department of Statistics, University of California Berkeley Technical Report, No.421 (1994)

- [8] Yoav Freund, Robert E. Schapire: "A decision-theoretic generalization of on-line learning and an application to boosting" Journal of Computer and System Sciences, Vol.55, No.1,pp.119-139 (1997)
- [9] T. Tsumura, F. Saitoh, S. Ishizu: "Knowledge Extraction from Textual Patent Data Using a Random Forest", J Jpn Ind Manage Assoc, Vol. 68, No. 3 pp. 161-170 (2017)
- [10] 柿崎 恒美, i-Construction : 建設現場の生産性向上について, 月刊建設, Vol.60, No.1 pp.6-9 (2016)
- [11] 近藤弘嗣: ICT 施工への取り組みについて; 建設機械 Vol.53, No.3, pp.33-38 (2017)
- [12] 藤野健一, 茂木正晴, 大槻崇: 建設現場における IT・RT 普及-次世代の施工技術に向けて; 土木技術資料 Vol.52, No.1, pp.16-19 (2010)
- [13] 原田純仁: ICT ブルドーザと ICT 油圧ショベルの開発; 計測と制御 Vol.55.No.6, pp.523-526 (2016)
- [14] 森川勝: 油圧ショベル・マシンガイダンスシステムについて: ICT 施工が現場を変える. はじめるならホルナビ; 建設機械, Vol.53, No.3, pp.62-65 (2017)
- [15] Y.Sakaida, D.Chugo, H.Yamamot, H.Asama: "The Analysis of Excavator Operation by Skillful Operator", Proc. of SICE Annual Conference, tokyo, pp.538-542 (2008)

- [16] 藤田光伸, 鎌田実, 宮田圭介: 機械操作における技量の抽出に関する研究 (第1報, 視覚情報からの認知能力の抽出); 日本機械学会論文集 C 編, Vol.67, No.653, pp.201-208, (2001)
- [17] 藤田光伸, 鎌田実, 宮田圭介: 機械操作における技量の抽出に関する研究 (第2報, 操作時の認知能力の解明および操作支援への応用手法の提案); 日本機械学会論文集 C 編, Vol.68, No.666, pp.547-554, (2002)
- [18] K. Koiwai, Y. Liao, T. Yamamoto, T. Nanjo, Y. Yamazaki, and Y. Fujimoto: "Feature extraction for excavator operation skill using CMAC", Journal of Robotics and Mechatronics , Vol.28, No.5 pp.715-721 (2016)
- [19] 境田右軌, 中後大輔, 川端邦明, 淺間一: 熟練オペレータによる油圧ショベルの掘削作業解析; ロボティクス・メカトロニクス講演会 2007 (ROBOMECH07) , 秋田, pp.1P1M07(1)-1P1M07(2), (2007).
- [20] Hiromu Imaji, Kazushige Koiwai, Toru Yamamoto, Koji Ueda, Yoichiro Yamazaki, Yoshiaki Fujimoto: "Human Skill Quantification for Excavator Operation using Random Forest", Proc. of ICAROB 2017, Miyazaki, pp.437-440, (2017)
- [21] 今地大武, 小岩井一茂, 山本透, 上田浩司, 山?洋一郎: ランダムフォレストを用いた油圧ショベルの操作スキルの定量化と評価; 第4回計測

- 自動制御学会制御部門マルチシンポジウム論文集, 岡山, pp.576-579, (2017)
- [22] 小岩井一茂, 今地大武, 山本透, 上田浩司, 山?洋一郎: 機械学習を用いた油圧ショベルの掘削操作における特徴抽出; 第61回システム制御情報学会研究発表講演会 (SCI'17), 京都, 261-1, (2017)
- [23] J. Rasmussen: "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-13, No.3, (1983)
- [24] K. Koiwai, Y. Liao, T. Yamamoto, T. Nanjo, Y. Yamazaki, and Y. Fujimoto: "Feature extraction for excavator operation skill using CMAC", Journal of Robotics and Mechatronics, Vol.28, No.5 pp.715-721 (2016)
- [25] Y. Kawazoe, K. Enomoto, J. Matsumoto, S. Okabe, and S. Handa: "Acquisition of Human Operator's Skills and Dexterity using Neural Network: Automatic Generation of Neural Controller from Chaotic Time Series Data during Stabilizing Control of an Inverted Pendulum", Proc. of 29th SICE Symposium on Intelligent Systems, pp. 217-222 (2002)
- [26] J. Kim, J. Lee, and Y. Lee: "Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree",

Healthcare Informatics Research, Vol.21, No.3 pp.167-174 (2015)

- [27] D.T.Pham and X.Liu: "Identification of linear and nonlinear dynamic systems using recurrent neural networks", *Artificial Intelligence in Engineering*, Vol. 8, No. 1, pp. 67-75 (1993)
- [28] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P-Y. Glorennec, H. Hjalmarsson and A. Jurditsky: "Nonlinear Black-box Modeling in System Identification: A Unified Overview", *Automatica*, Vol 31, No 12, pp. 1691-1724 (1995)
- [29] 牛田 俊, 木村 英紀:「Just-In-Time モデリング技術を用いた非線形システムの同定と制御」, *計測と制御*, Vol. 44, No. 2, pp. 102-106 (2005)
- [30] A. Stenman, F. Gustafsson and L. Ljung: "Just in Time Models for Dynamical Systems", *Proceedings of 35th IEEE Conference on Decision and Control*, pp. 3-6 (1996)
- [31] 岸 真友, 木村 和喜, 山本 茂:「Just-In-Time モデリング技術による鉄鋼製品の収縮率予測」, *計測と制御*, Vol. 44, No. 2, pp. 116-119 (2005)
- [32] T. Yamamoto, K. Takao and T. Yamada: "Design of a Data-Driven PID Controller", *IEEE Trans on Control Systems Technology*, Vol. 17, No. 1, pp. 29-39 (2009)

- [33] 高尾 健司, 山本 透, 雛元 孝夫: 「Memory-Based 型システム同定による一般化予測制御系の一設計」, 電気学会論文誌 C, Vol. 125, No. 3 pp. 442-449 (2005)
- [34] 堤 健人, 山本 透: 「データ駆動型内部モデル制御系の一設計」, 電気学会論文誌 C, Vol. 131, No. 4 pp. 781-787 (2011)
- [35] 黒木 学, 宮川 雅巳, 川田 亮平: 「条件付き操作変数法の推定精度と操作変数の選択」, 応用統計学, Vol. 32, No. 2 pp. 89-100 (2003)
- [36] 和田 拓也, 藤原 幸一, 加納 学: 「非線形システムの時系列データを対象とした因果推論手法の比較」, "Comparison of causal inference methods for time series data of nonlinear systems", 計測自動制御学会制御部門マルチシンポジウム (2018)
- [37] D. E. Farrar, R. R. Glauber: "Multicollinearity in Regression Analysis: The Problem Revisited", The Review of Economics and Statistics, Vol. 49, No. 1, pp. 92-107 (1967)
- [38] 津村 拓海, 斎藤 史哲, 石津 昌平: 「ランダムフォレストを用いた特許に関する文書データからの技術適用領域に関する知識選択」, 日本経営工学会論文誌, Vol. 68, No. 3 pp. 161-170 (2017)
- [39] X. Wang, X. Liu, W. Pedryc, and L. Zhang: "Fuzzy rule based decision trees", Pattern Recognition, Vol.48, No.1 pp.50-59 (2015)

本論文に関連する掲載論文

- [1] 小岩井 一茂, 今地 大武, 山本 透, 上田 浩司, 山崎 洋一郎: 機械学習を用いた油圧ショベルの掘削操作における特徴分析法の一提案; システム制御情報学会論文誌 31(9) 328-335 2018/9
- [2] 今地 大武, 木下 拓也, 山本 透, 伊藤 啓介, 吉田 昌弘, 中川 正: ランダムフォレストによる変数選択に基づくデータベース駆動型モデリングとその線形空燃比センサ出力予測への応用; 電気学会論文誌C (電子・情報・システム部門誌) 139(8) 2019/8
- [3] Hiromu Imaji, Takuya Kinoshita, Toru Yamamoto: "Design of a Database-Driven Modeling based on Variable Selection using a Random Forest", IEEE International Conference on System, Man and Cybernetics (IEEE SMC 2018) 2018/10/8