レンタル用モジュラー建造物の返却時期予測におけるデータ前処理手法の提案
Data Preprocessing Method in Predicting Return Date of Rental Unit Houses

舟久保 龍成
FUNAKUBO, Ryusei

概要：三協フロンテア株式会社ではユニットハウスのレンタル事業を行っている．ユニットハウスとは再利用可能な組み立て式の建築物で，仮設住宅や仮設事務所等に使用されている．貸出し時にユニットハウスの返却予定日は申告されるものの，状況に応じて使用期間を借り手が変更することができる．ユニットハウスの生産は在庫数によって決める必要があるが，返却予定日が変更される可能性があるため在庫を予測して生産計画を立てることが難しい．2015 年より決定木を用いて返却時期の予測をし，2018 年より決定木と遺伝的アルゴリズムを用いて返却時期の予測をしたが，高い予測精度が得られていない．先行研究に使用したデータを調査した結果，データの属性の調整，および属性値の変換に問題があることがわかった．本研究では，精度の良い決定木の生成を目的として，共同共進化に基づく適切なデータ前処理手法を提案する．

Summary: SANKYOFRONTIER CO., LTD does rental business of rental modular buildings, that are called “unit houses.” Unit house is a reusable assembled building, used as temporary housing, temporary office, etc. The expected date to return unit houses is declared when a customer rents them. However, it is possible for the customer to change the return date according to the situation. Therefore, it is difficult to predict number of stocks and make a production schedule. We started research on predicting return date using a decision tree in 2015 and predicting return date using a decision tree and Genetic Algorithm in 2018. However, high accuracy has not been obtained. As a result of investigating the data used in the previous research, it was found that there were problems in adjusting the attributes of the data and converting the attribute values. In this research, I propose an appropriate data preprocessing method using on Cooperative Coevolution for the purpose of generating accurate decision trees.

キーワード: データマイニング・決定木・遺伝的アルゴリズム・共同共進化・C4.5
Keywords: Data mining, Decision tree, Genetic algorithm, Cooperative Coevolution, C4.5

1. はじめに

1.1. 研究の背景

現在，三協フロンテア株式会社ではユニットハウスのレンタル事業を行っている．ユニットハウスとは再利用可能な組み立て式の建築物で，仮設住宅や仮設事務所等に使用されている．貸出し時にユニットハウスの返却予定日は申告されるものの，状況に応じて使用期間を借り手が変更することができる．ユニットハウスの生産は在庫数によって決める必要があるが，返却予定日が変更される可能性があるため在庫を予測して生産計画を立てることが難しい．データ分析に利用するために，ユニットハウスのレンタル事業に関するデータを記録しているが，データ分析の知識を持っている社員が少なく，データを有効活用できていない．2015 年より，決定木を用いた返却時期の予測が行われているが，高い予測精度は得られていない．決定木とはツリー構造で表された分類規則であり，分類基準が明確に示されるためデータ分析に詳しくない人でも，各クラスを特徴づける要素を直感的に知ることができる点が長所である．ユニットハウスの返却時期予測においては，予測結果だけでなく返却時期を左右する要因を知ることも重要であるため，決定木の使用が有効であると考えられる．

1.2. 先行研究

先行研究として，2018 年より決定木と遺伝的アルゴリズムを用いた返却時期予測の研究[1]が進められている．決定木で高い予測精度が得られていない原因として，ノイズデータの存在を挙げた．ノイズデータとは例外的なデータのことであり，データ分析の前に対象データから除去することで分析精度の向上が見込まれる．予測精度の高い決定木の生成を目的として，遺伝的アルゴリズムを用いて実データに含まれるノイズデータを選別した．先行研究の結果，生成された決定木の正解率や可読性を向上することができたが，実用に耐えうる十分な結果は得られなかった．

決定木生成に使用したデータセットを調査した結果，データセットを構成する属性の選択，および属性値のカテゴリ表現に問題があることがわかった．決定木生成アルゴリズムには，C4.5 を採用している．C4.5 では，属性

この研究の一部は，情報処理学会第 82 回全国大会において発表の予定である．

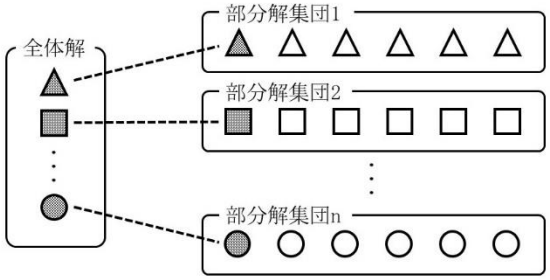

図 1 共同共進化における全体解の例

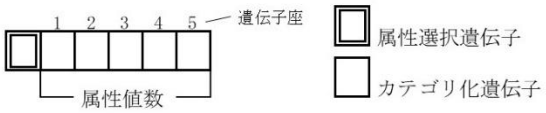

図 2 部分解の染色体の構成

図 3 部分解の適用例

値の種類数により各ノードの分岐数が決まるため，属性値をカテゴリ化して種類数を減らす必要がある．また．決定木の精度を上げるには重要な属性を組み合わせてデータセットを作成する必要がある．決定木生成に使用するデータセットの属性の組合せと属性値のカテゴリ表現を適切に設定することで，予測精度の向上が見込める．

1.3. 研究の目的

本研究では，正解率が高く，可読性の高い決定木の生成を目的として，共同共進化[2]に基づく適切なデータ前処理手法を提案する．属性選択と属性値のカテゴリ化をデータ前処理の対象とし，期限通り返却されるか，期限よりも遅れるか，期限よりも早まるかの 3 クラスに分類する決定木を生成して返却時期を予測する．

2. 共同共進化

共同共進化は，複雑で動的な問題を扱うことを目的として提案された遺伝的アルゴリズムである．統治分割法に基づき，問題に対する解がいくつかの部分の組み合わせにより構成されると考える．各部分に対する解の候補を個体として保存し，各個体を協力的な相互作用によって進化させることで最適解を構成するために必要な部分を獲得する．解の一部を部分解，部分解の組合せによって形成される完全な解を全体解と呼ぶ．共同共進化における全体解の一例を図 1 に示す．

共同共進化では，部分解を個体とする集団と部分解の組合せを個体とする全体解集団の両集団を並行して進化させる．部分解集団の各個体は，解の一部の最適化と，他の部分を最適化する個体と協働して解全体の最適化という 2 つの目標をもって進化する．前者の対象となる部分は個体ごとに異なるため，異なる目標を持った複数の探索が並行して行われることになる．また，後者のために個体の適応度は全体解の評価に基づいて決定される．

3. 共生共進化に基づくデータの前処理手法

3.1. 提案手法

本手法では，共同共進化を用いて決定木生成に適したデータセットの属性と属性値のカテゴリ表現を探索する．データの属性が A_1～A_N であるとき，各属性に対応する N 個の部分解集団 S_1～S_N と 1 個の全体解集団を並行進化させる．提案手法では，訓練データを決定木生成用データと，精度測定用データにわけ，前者を C4.5 の訓練データ，後者を共同共進化の適応度計算に使用する．

3.2. 全体解

データに含まれる属性数を N とすると，全体解は N 個の部分解へのポインタで構成される．i 番目のポインタは部分解集団 S_i の個体を指す．1 個体で決定木生成に適したデータの前処理を表現する．

3.3. 部分解

部分解の染色体の構成を図 2 に示す．部分解集団 S_i に所属する部分解は，属性 A_i 使用するか否か，および属性 A_i の属性値のカテゴリ化を表す．属性 A_i の属性値の種類数を $n(A_i)$とすると，部分解の染色体は，1 個の属性選択遺伝子と，$n(A_i)$個のカテゴリ化遺伝子からなる．属性選択遺伝子は 0 または 1 の値を取り，0 のときデータセットに属性 A_i を使用，1 のとき使用しないことを表す．カテゴリ化遺伝子はそれぞれ 1 つの属性値に紐づけられており，0 以上 $n(A_i)$未満の整数を取る．遺伝子が同じ値の場合，紐づく属性値は同一カテゴリに属すとみなす．

属性「数量」の属性値が 1 個~5 個の場合の部分解の表現例を図 3 に示す．属性選択遺伝子は 0 であり，属性「数量」はデータセットに使用する．属性値はカテゴリ化遺伝子を元に，「1 個，3 個，5 個」と「2 個，4 個」の 2 カテゴリに分けられる．

3.4. 適応度関数

各全体解を評価するために，全体解が表す方法で前処理された決定木生成用データを用いて C4.5 により決定木を生成し，精度測定用データにより評価する．得られた決定木 t の正解率を $x(t)$，ノード数を $y(t)$，クラスごとの正解率の積を $z(t)$として適応度 $f(t)$を式(1)より算出する．

$$f(t) = x(t) \times \frac{1}{1 + e^{-(-1*y(t)+1000)*0.01}} \times z(t) \qquad (1)$$

クラスごとの正解率の積 $z(t)$は,クラス i の正解率を $z_i(t)$として式(2)より算出する．

$$z(t) = \prod z_i(t) \qquad (2)$$

また，$z_i(t)$が極端に小さい値にならないように，事前に前処理をしていない決定木生成用データで決定木 t'を生成し，精度測定用データで $z_i(t')$の数を求めた．$z_i(t')$が 0.3 未満の場合，式 2 における $z_i(t)$に対して 1.0 を加算する．例外的に z_i が極端に小さい値になるのを防ぐことで，探索が効果的に進むようにする．

部分解の適応度は，所属する全体解の適応度をそのま

表 1 各データのデータ数とクラスごとのデータ数

	総数	期限通り	遅れる	早まる
決定木生成	11933	8566	3196	171
精度測定	2614	1335	1222	57
テスト	1244	518	646	80

表 2 データに含まれる属性

属性
出荷月，返却予定月，販売タイプ，住所，数量， 使用用途，クライアントの業種，現場区分， 出庫されたセンターの住所，担当営業コード

ま反映させる．1 つの部分解が複数の全体解に所属している場合は，最も適応度の値が良いものを反映する．

3.5. 世代交代モデル

部分解の世代交代は HS モデル[3]により行う．HS モデルの処理の手順を以下に示す．

Step1. 上位半数をそのまま次世代に残す．

Step2. 上位 25%のある個体のなかから 2 個体をランダムに選択し，親個体とする．

Step3. 交叉により生成された 2 つの新しい個体のいずれかと，親個体のいずれかのコピーを子個体とする．子個体に対して，ある確率で突然変異を発生させる．

Step4. 評価の低い 2 つの個体と 2 つの子個体を置き換える．

Step5. 上位 25%のすべての個体に対して Step2~Step4 を繰り返す．

Step2 の突然変異において部分解集団では，遺伝子の値を 0 以上 $n(A_i)$未満の整数でランダムに書き換える．

全体解集団の世代交代では，親の選出にランキング選択，交叉方法に一様交叉を採用する．また，突然変異は以下の 2 種類を行う．

Mutation1. 別の部分解へのポインタに置き換える．

Mutation2. 50%程度の確率で，部分解集団の親個体だった部分解へのポインタを，該当部分解から生成された子個体へのポインタに書き換える．

Step1~Step4 は非常に強い収束力をもつ戦略であるが，Mutation2 によって集団の多様性が維持される．また，交叉で生成された子個体への探索が可能になるだけでなく，親個体のコピーへのポインタによって部分解集団における有害な突然変異から迅速に復旧することができる．

3.6. 部分解集団の遺伝子の修正

部分解の個体は，0 以上 $n(A_i)$未満の値にしてカテゴリ化をしているため，交叉や突然変異によって次世代個体に含まれる属性値の値が 0 からの連番にならなくなる問題や，属性値が 1 種類になる問題が発生する．部分解集団の世代交代では，次世代個体が生成されたタイミングで遺伝子に不備がある個体を修正する．

3.7. 最適解の探索

提案手法の処理の流れを下に示す．

Step1. 全体解と各部分解の初期集団の個体を生成する

表 3 決定木の精度

	決定木 A	決定木 B
正解率(%)	53.6	41.2
ノード数	157	6867
木の高さ	4	7

表 4 各データセットに含まれる属性

データセット A	データセット B
住所，使用用途， クライアントの業種， 現場区分	出荷月，返却予定月，販売 タイプ，住所，数量，使用 用途，クライアントの業種

Step2. 全体解を元にデータの前処理をする．

Step3. 決定木を生成し全体解と全体解に紐づく部分解の適応度を求める．

Step4. 適応度に基づいて全体解集団と各部分解集団の世代交代をする．

Step5. 部分解集団において，世代交代によって生成された個体に不備があった場合，該当個体の遺伝子を修正する．

Step6. Step2 から Step5 までを任意の回数繰り返し，最終世代で最良個体を出力する．

4. 評価実験

4.1. データセットと共同共進化のパラメータ

三協フロンテア株式会社において貸出されたユニットハウスの返却時のデータを使用して提案手法を評価した．データの記録期間は 2011 年度から 2017 年度までの 7 年間であり，2011 年度から 2015 年度までのデータを決定木生成用データ，2016 年度のデータを精度測定用データ，2017 年度のデータをテストデータとする．各データを提案手法で前処理したものをデータセット A，人手で前処理したものをデータセット B とする．各データのデータ数とクラスごとのデータ数を表 1 に，本研究で使用するデータに含まれる属性を表 2 に示す．

提案手法の共同共進化のパラメータは，突然変異確率 1%，全体解の集団サイズを 400，部分解の集団サイズを 400，世代交代数を 500 回とした．

4.2. 生成された決定木の評価

得られた決定木の正解率とノード数を表 3 に示す．データセット A で生成した決定木の正解率はデータセット B で生成した決定木の正解率より約 12%高かった．ノード数は 6867 から 157 に削減することができ，十分な可読性が得られたといえる．提案手法では，各個体の適応度の算出に精度測定用データを毎回使用しているため，過適合が懸念されるが，評価実験では，比較したデータセット B の決定木よりも高い正解率を獲得することができているため汎化性能も保たれると考えられる．しかし，実用的な正解率とはいえないため，さらなる正解率の向上のための改善が望まれる．

4.3. 特徴選択の評価

表 4 に各データセットに含まれる属性を示す．提案手

表 5 カテゴリ化による属性値の種類数の変化

属性	元データ	処理後
住所	7	3
クライアントの業種	7	4
使用用途	5	2
現場区分	5	4

表 6 属性「住所」のカテゴリ化の結果

カテゴリ A	カテゴリ B	カテゴリ C
宮城	東京, 神奈川, 埼玉	茨城, 千葉, 愛知

法の特徴選択の結果，データセット A の属性は 4 種類になった．提案手法によって選出された 4 属性のうち「使用用途」，「クライアントの業種」，「現場区分」の 3 つの属性は内容がクライアントに関わる属性であることから，返却時期はクライアントに強く依存していると考えられる．属性選択によって新たな属性の組み合わせを見つけることができたが，得られた決定木の正解率が実用的な値ではないことから，現在記録しているユニットハウスのレンタル事業のデータでは，十分な正解率を得るために必要な属性が不足していると考えられる．例えば，クライアントの規模や，使用用途の規模，動員数，予算などといった情報を分析に使うことができるならば，より重要な属性を見つけ出し，決定木の正解率を向上させることができると考える．

4.4. 属性値のカテゴリ化の評価

カテゴリ化による属性値の種類数の変化を表 5 に，属性「住所」のカテゴリ化の結果を表 6 に示す．提案手法のカテゴリ化によって，属性値の種類数の最大値は 7 個から 4 個に減少した．決定木の可読性を上げるように適応度関数を設定したため，必然的に属性値の種類数は減ったが，決定木の正解率が向上しているため，効果的に同じ構造を持った属性値をカテゴリ化できたと考えられる．属性「住所」では，近い地域でカテゴリ分けされているように考えられる．しかし，カテゴリ C の「茨城県，千葉県，愛知県」では，かなり離れた地域にある愛知県が茨城県，千葉県と同カテゴリに分類されている．提案手法のカテゴリ化によって，3 つの地域には位置情報以外の共通点がある可能性が示唆されたといえる．しかし，どのような共通点があるかは把握できないため，事業内容に詳しい人と精査することで新たな知見が得られ，さらなる精度向上のきっかけが得られるかもしれない．

4.5. 提案手法の共同共進化の設定

正解率とノード数の推移を図 4 に示す．正解率とノード数が互いに増減しながら収束しており，両方を考慮した探索ができているといえる．しかし，属性選択と属性値のカテゴリ化を同時にする大規模な探索にもかかわらず 100 世代目には収束している．世代交代に HS モデルを採用することによって，通常の世代交代よりも個体の多様性を維持できているが，今以上に多様性を維持した

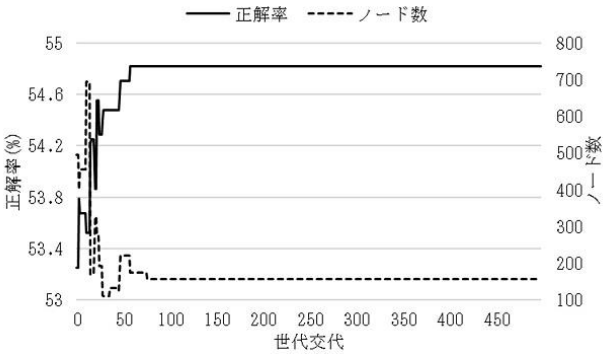

図 4 正解率とノード数の推移

進化戦略を考える必要がある．

5. おわりに

本研究では，正解率が高く，可読性の高い決定木の生成を目的として，共同共進化に基づく適切なデータ前処理手法を提案した．評価実験の結果では，人手によるデータの前処理をしたデータセットによって得られた決定木よりも正解率，可読性ともに改善されたため本手法の有用性を示せた．可読性においては，木の構造を容易に把握することができるようになったため，十分な成果といえる．しかし，正解率を実用レベルにまで改善することはできなかったため手法の見直しが必要である．本手法は，機械的なカテゴリ化をしているため，出力結果の調査に，データに関するより深い知識が必要となる．

本研究の特色は，実データを利用している点にもある．今回扱ったユニットハウスのレンタル事業のデータは，複数の担当者によって手入力された複数のデータを継ぎ接ぎして作成されたデータである．データの整合性が取れない問題や，担当者による入力項目の誤認識や入力不備などが起きている問題があった．実データでは，ヒューマンエラーが大きなノイズとなり得るため，入手したデータがどのように入力，保管，加工されたかを確かめる必要がある．また，データが不足すると分析手法が限られるだけでなく，十分な精度の分析結果が得られなくなる恐れがあるため，どのように分析に使うかを想定してデータを記録することも重要となる．実データの分析では，データを準備することが分析をすることよりも重要となるが，データの不備を取り除いたとしても，容易に良い分析結果が得られるわけではないため，忍耐強く試行錯誤する必要がある．

参考文献

[1] 舟久保龍成, 土屋直樹, スティーヴェンクレイネス, 大谷紀子, "モジュラー建築ユニットの返却時期を予測する決定木の GA による精度向上", 情報処理学会第 80 回全国大会予稿集, Vol.1, pp.299-300, 2018.

[2] Potter, M.A. and De Jong, K.A., "A Cooperative Coevolutionary Approach to Function Optimization", in Proceedings of the 3rd Parallel Problem Solving from Nature, pp.249-257, 1994.

[3] Moriarty, D.E. and Miikkulainen, R., "Efficient Reinforcement Learning through Symbiotic Evolution", Machine Learning, Vol. 22, pp.11-32, 1996.