

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

舟久保龍成† 土屋直樹‡ スティーヴン・クレイネス§ 大谷紀子†

東京都市大学メディア情報学部† 三協フロンテア株式会社‡ 東京都市大学共通教育部§

1. 研究の背景と目的

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ているが、返却予定日が変更す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た返却時期予測の研究が進められている。決定木とはツリー構造で表された分類規則であり、分類基準が明確に記されるため、各クラスを特徴づける要素を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が重要であるため、決定木の使用が有効であると考えられる。返却時期が予定日より早まるか、遅くなるか、予定通りに返却されるかを判別する決定木の生成をC4.5や共生進化に基づく手法[1]で試みているが、実用に耐えうる十分な予測精度は得られていない。原因の1つにノイズデータの存在が考えられる。ノイズデータとは例外的なデータであり、データ分析の前に対象データから除去することで精度の向上が望める。本研究では、予測精度の高い決定木を生成することを目的として、実データからノイズデータを選別する手法を提案する。

2. ノイズデータの選別方法

本手法では、遺伝的アルゴリズム(GA; Genetic Algorithm)を用いる。GAとは生物が進化する過程を模倣した最適解探索アルゴリズムである。問題に対する解を染色体で表現し、解

としての良さを適応度として、複数個体の中から適応度の高い個体の形質が継承されるように、選択、交叉、突然変異により次世代の個体集団を生成する。世代交代を繰り返し、適応度の高い個体を探索する。終了条件は目的に応じて設定する。

本手法では、GAの染色体の長さをデータセットに含まれるデータ数とし、染色体の遺伝子はデータセットの各データと1対1で対応させる。遺伝子の値は0と1で表現し、0の場合ノイズでないデータ、1の場合ノイズデータとする。10%の確率で各遺伝子が1になるようにして初期集団の個体を n 個体生成する。また、世代交代時にノイズデータの割合が m を超えた個体は新たに生成した個体と置き換える。各個体を評価するために、染色体が示すデータから決定木生成アルゴリズムC4.5を用いて決定木を生成する。使用したデータ数 x 、生成された決定木の正解率 y の個体の適応度 $f(x, y)$ は式(1)により算出する。定数 α はシグモイド関数のゲインを表し、保持したいデータ数に応じて設定する。世代交代する際の親個体の選択方法はランキング選択を使用し、一様交叉と、0.1%の突然変異によって子の個体を生成する。

$$f(x, y) = \left(y + \frac{1}{1 + e^{-\alpha x}} \right)^3 \quad (1)$$

データ数が極端に少ないことで決定木の正解率が高くなる可能性があるため、一定のデータ数を保持したまま、決定木の正解率が高くなるように適応度関数を設定し、出力結果が目的に沿うようにする。世代交代を i 回繰り返し、適応度の一番高い個体を出力する。

使用するデータのクラスは、 $-1, 0, 1$ で表現されており、それぞれ返却予定日より早く返却される、予定通り返却される、返却予定日より遅く返却される、を表す。

3. 評価実験

三協フロンテアにおけるユニットハウスの返却時の7079データセットのうち、6372を訓練デ

Using a GA to Improve Decision Trees for Return Time Prediction of Rental Modular Buildings

† Ryusei Funakubo, Noriko Otani, Faculty of Informatics, Tokyo City University

‡ Naoki Tsuchiya, SANKYOFROntier CO.,LTD

§ Steven Kraines, Faculty of Liberal Arts, Tokyo City University

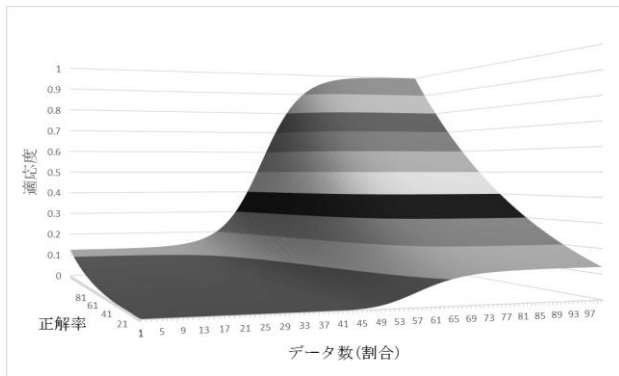


図 1：評価実験で使った適応度関数

ータ, 707 をテストデータとして評価実験を行った. GA の各パラメータは, 個体数 n を 1000, 世代交代数 i を 1000, ノイズデータの最大含有割合 m を 0.2 とした. また, 決定木の生成に使用するデータ数が 9 割を下回った場合, 適応度への影響が徐々に大きくなるように定数 α を $1/6$ とした. 評価実験で使った適応度関数のグラフを図 1 に示す. データ数は割合で表してある.

データの属性は, 出荷月, 返却月, 販売タイプ, 住所, 数量, 用途, 業種である. 提案手法を用いてノイズを除去したデータセット, オリジナルのデータセット, 各データの選択確率を 90%として生成したデータセットのデータ数, および各データセットで C4.5 により生成した決定木の正解率, 木の高さ, ノード数を表 1 に, 提案手法を用いて生成した決定木とオリジナルデータを用いて生成した決定木の根ノードとレベル 1 のノードの属性を表 2 に, GA の世代ごとの最適個体の値を図 2 に示す. 初動ではデータ数が減ることで正解率が上がっているが, 世代交代が進むにつれてデータ数が上がっている. ランダム抽出のデータセットの結果としては, 用意した 100 個のデータセットに関する平均と標準偏差を示す.

提案手法を用いて生成した決定木の根ノードの属性には, 現場, イベント, 土地建築販売, 選挙, その他の 5 つの属性値がある. 現場と選挙の子ノードには返却月の属性, イベントの子ノードには住所の属性, 土地建築販売の子ノードには業種の属性がある.

4. 考察

評価実験の結果, GA によってノイズと判定されたデータを取除くことで, オリジナルのデータで生成した決定木より正解率を約 7%上げることができた. しかし, 正解率は 49.08%と実用的ではないため, 改善の必要がある.

予測方法に決定木を用いることで, 普段からデータ分析などに関わらない人にも, 予測結果

表 1：各データセットで生成された決定木の性質

データセット	提案手法	オリジナル	ランダム抽出
データ数	5192	6372	5656.28±243.31
正解率	49.08%	42.15%	41.53%±0.81
木の高さ	7	7	7±0
ノード数	1865	6013	5604.73±185.86

表 2：ノードの属性

データセット	提案手法	オリジナル
根ノード	用途	用途
レベル 1	返却月, 住所, 業種	返却月, 住所, 業種, 数量

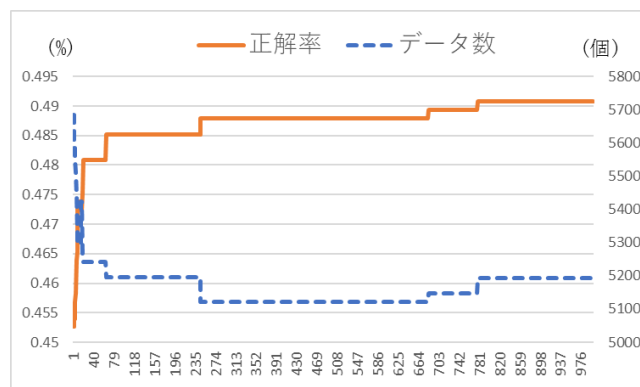


図 2：評価実験で使った適応度関数

と予測の要因を直感的に理解させる狙いがある. オリジナルのデータから生成した決定木ではノードが 6013 個あり, 出力される決定木が大きすぎるため可読性が低い. 提案手法によって少ないノード数の決定木が得られたことで, 可読性を高めることができた. しかし, まだノード数が 2000 近くあるため, 可読性の高い決定木とはいえない. 表 2 より, 根ノードの属性が変化していないことから, 用途が重要な属性であることがわかる. また, レベル 1 のノードの属性は変化していることから, 新たに重要である属性を特定できたと考えられる.

提案手法を用いて生成した決定木より, 属性値に現場と選挙を含むデータは, 返却される時期に影響を受けていることがわかる. イベントを含むデータは, ユニットハウスが使用される地域や, 出庫されるセンターからの距離に影響を受けていると考えられる. 土地建築販売の子ノードである業種には建築業と不動産業の属性があり, 業務形態の違いが返却に影響していると考えられる.

参考文献

[1] 大谷紀子, 志村正道, "共生進化に基づく簡素な決定木の生成", 人工知能学会論文誌, Vol.19, No.5, pp.399-404, 2004.