機械学習による賃貸物件データを用いた家賃推定手法の検討

Examination of estimation methods of house rent by realestate information and machine learning

加藤 暢之

Nobuyuki Kato

岡山大学 太田研究室

Ohta Laboratory, Okayama University

概要 本研究では賃貸物件データを利用した賃料の推定を機械学習により行う．学習データは国立情報学研究所の提供する賃貸物件データを扱った LIFULL HOME'S データセットを使用する．利用する学習モデルは重回帰分析，ラッソ回帰，リッジ回帰，サポートベクター回帰の 4 つである．精度比較は推定した価格の誤差から算出した決定係数で行う．

1 はじめに

近年 Web からアクセスできる情報の量は加速度的に増加している．不動産情報もその一つであり，不動産仲介業者のもつ情報が不動産・住宅情報サイトに集約されている．本研究では不動産・住宅情報サイト LIFULL HOME'S に掲載されたデータのスナップショットである LIFULL HOME'S データセット[*1]の賃貸物件データに注目する．LIFULL HOME'S は不動産・住宅情報サイトの一つであり，産経広告社の調べ[*2]によると 2018 年 1 月 7 日の時点で賃貸，新築，中古マンション，新築戸建，中古戸建の 5 カテゴリ物件総数が 7,694,046 件で他の不動産・住宅情報サイトと比べると物件総数はトップである．

2 データセット

本研究では国立情報学研究所が提供している LIFULL HOME'S データセットを利用する．LIFULL HOME'S データセットは不動産・住宅情報サイト LIFULL HOME'S に 2015 年 9 月時点で掲載されていた賃貸物件データのスナップショットである．データ内容は 2 種類あり，賃貸物件データと賃貸物件の画像データであるが，本研究では賃貸物件データのみを扱う．賃貸物件データは全国約 533 万件の物件における賃料，面積，立地等の情報を含んでいる．

3 前処理

LIFULL HOME'S データセットは 533 万件の物件に対してそれぞれ 71 項目の情報を記述しているデータセットである．本研究では岡山県のデータ 78,019 件を対象として家賃を推定する実験を行った．また，71 種類の特徴から物件価格に影響を与える可能性のあるものを人手で 24 種類選択した．選択した項目とデータの欠損数を表 1 に示す．

表 1: データの持つ特徴の項目と欠損数 (78,019 件中)

項目	欠損数	項目	欠損数
物件種別	0	建物階数	101
総戸数	63,516	築年数	29
都道府県	0	部屋階数	890
市区町村	0	向き	5,531
路線 1	108	間取り部屋数	60
徒歩距離 1	108	間取り部屋種類	102
路線 2	20,639	賃料	0
徒歩距離 2	20,639	小学校距離	31,239
用途地域	75,400	中学校距離	34,808
都市計画	75,418	コンビニ距離	20,305
建物構造	220	スーパー距離	22,861
建物面積	28	総合病院距離	33,901

表 1 より，専門的な項目や距離を示す項目はデータが欠損している割合が高い．そのため欠損データを除外するとデータの絶対数が減少するため，欠損データを除外しない前処理をした．連続値に対しては欠損値以外から平均値を算出し補填した後，標準化を行った．また連続値以外の文字列で表現された項目に対しては one-hot エンコーディングによる変換を行った．

4 学習モデル

家賃の推定には以下のモデルを使用した．

重回帰分析 実測値と推定値の平均二乗誤差を損失関数として回帰的に各項目の重みを決定する．重回帰分析モデル [1] をベースラインとして扱う．

ラッソ回帰 基本は重回帰分析と同様であるが，L1 正則化項 (絶対値の和) を損失関数に付加することで過学習を抑制する．これによりいくつかの重みは 0 になりパラメータを圧縮することができる．ラッソ回帰は項目数が多く，重要なものが僅かしかない場合に有効な手法である．

リッジ回帰 ラッソ回帰の正則化項を L2 正則化項 (二乗和) に変更した手法である．ラッソ回帰と違い重みが 0 になることはないため，パラメータの圧縮が望ましくない場合に用いる．

[*1] https://www.nii.ac.jp/dsc/idr/lifull/homes.html

[*2] https://www.sankeibiz.jp/business/news/180130/bsd1801300500005-n1.htm

表 2: 各モデルの決定係数 R^2

モデル	訓練データ	検証データ	テストデータ
重回帰分析	0.7845	-2.4731	-1.4035
ラッソ回帰	0.5245	0.4891	0.5058
リッジ回帰	0.7469	0.4734	0.6260
SVR	0.8902	0.8246	0.8413

サポートベクター回帰　サポートベクター回帰 (support vector regression) (以下 SVR) は誤差と同時に重みも最小化することで過学習を防ぐ手法である．本実験ではガウシアンカーネルを使った SVR を使用する．ハイパーパラメータである C，ε，γ の 3 つについてグリッドサーチで最適値を算出すると計算時間が膨大になるため，簡略化のためアルゴリズム [2] を使用した．

5　評価実験

5.1　モデル精度の比較

本節ではモデルの精度を決定係数 R^2 によって比較する．本実験ではデータを 10 分割し，そのうち 9 つを訓練データ，訓練データのうち 1 つを検証データ，残りの 1 つをテストデータとして評価する．決定係数 R^2 は式 (1) で表される．

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{1}$$

式 (1) において，y は実測値，$\hat{y}$ は推定値，$\bar{y}$ は実測値の平均，i はデータのインデックスである．各モデルにおいて訓練データ，検証データ，テストデータのそれぞれについて決定係数を算出した結果は表 2 となった．表 2 より重回帰分析モデルは訓練データ全体に対してはよく表現できているがテストデータでは精度が非常に低いため過学習が起きていると考えられる．ラッソ回帰は過学習は起きていないが全体的に精度が低い．リッジ回帰はラッソ回帰と比較して訓練データをよく表現しておりテストデータの推定精度も向上している．SVR は全データにおいて高い精度を出していることからテストデータでの決定係数も信用できる．

5.2　賃料の推定結果

5.1 で使用したモデルのうち，重回帰分析モデルの推定結果を図 1 に，SVR の推定結果を図 2 に示す．

図 1 と図 2 を比較すると，重回帰分析では推定の誤差が大きく，大きく実測値から外れたサンプルがあることが読み取れる．一方 SVR では，全体を通して比較的に推定誤差が小さく，価格が実測値の平均から大きく外れたサンプルに対してもある程度対応できていることが分かる．また実測値の曲線から，価格に対するクラスタリングにより極端な物件を別途学習すると精度が向上する可能性がある．

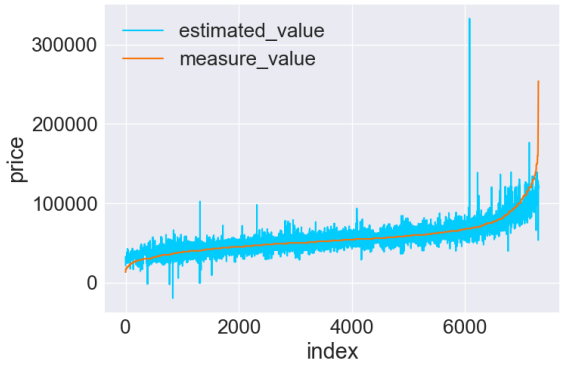

図 1: 重回帰分析

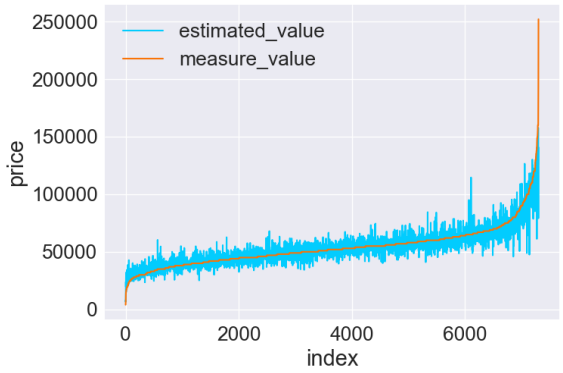

図 2: SVR

6　まとめ

本研究では，LIFULL HOME'S データセットを用いた賃料推定手法について検討した．本研究で最も良い精度を出したのはガウシアンカーネルを用いた SVR であったが SVR はデータ数が膨大になると精度が落ちる性質があるため，岡山県以外のデータも扱う場合については他の手法の検討も必要である．

参考文献

[1] 加藤暢之，新妻弘崇，太田学，“重回帰分析による土地価格推定の一手法”，第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018)，pp.1-6，2018.

[2] H. Kaneko, K. Funatsu, “Fast optimization of hyperparameters for support vector regression models with highly predictive ability”, Chemometrics and Intelligent Laboratory Systems, vol. 142, pp. 64–69, March 2015.