G-30

回帰分析とランダムフォレストを組み合わせた和牛の枝肉重量の予測手法

Predicting Carcass Weight of Wagyu with Regression Analysis and Random Forest

塚本 真伍[*1]　池上春香[*2]　松橋珠子[*2]　松本和也[*2]　吉廣卓哉[*3]
Shingo Tsukamoto Haruka Ikegami Tamako Matsuhashi Kazuya Matsumoto Takuya Yoshihiro

1　はじめに

和牛は海外からも注目されている日本の高級食材であり、和牛の価値を高めることに関心が寄せられている。和牛の経済的価値は様々な指標によって評価されるが、主な評価指標として枝肉重量、BMS(Beef Marbling Standard)、歩留基準値、バラの厚さ、皮下脂肪厚、ロース芯面積の 6 つの形質が挙げられる。その中でも特に枝肉重量は出荷される和牛の総量を決めるため特に重要視されている。

和牛の形質は遺伝的要因と環境要因の 2 種類の要因によって決まるとされている。繁殖農家は子牛を数カ月育成して出荷する。肥育農家は購入した子牛を肥育する際に 2 種類の要因のうち、環境要因を操作することでしか肉質を変化させることができない。環境要因は肉牛の成長に影響を及ぼす要因であり、肥育時の飼料や気候、運動量等多様な要素が影響を与える。肥育農家は環境要因を操作することで形質の向上を目指し日々様々な努力を行っている。しかし、各農家の肥育方法はこれまでの経験による知識やノウハウを基に行われており、肥育方法の移転や継承が困難であると同時に、安定した肉牛の生産を困難にしている。この問題を解決する科学的な根拠に基づいた、肉牛の生産性を向上させる肥育方法が求められている

本研究の目的は、肥育期間中に測定した和牛の導入時体重とタンパク質発現量から、屠殺時の枝肉重量の予測を行うことである。屠殺時の枝肉重量が肥育中に予測可能になることで、牛の出荷時期の見定めや、効率的な肥育方法を確立する手がかりが掴めると考えられる。

本研究では、回帰分析とランダムフォレスト法を組み合わせた、和牛の枝肉重量の予測手法を提案する。和牛の枝肉重量と導入時体重の間には大きい相関があることが知られている。この相関を活かした予測手法の構築を目指した。そこで、まずは相関によって説明できる部分については単回帰により予測し、単回帰で予測できない誤差にあたる部分を、分析データに機械学習を適用することにより補助することで、より精度の高い予測が可能であると考えた。

適用する機械学習法としては、ランダムフォレスト法を選んだ。説明変数となる分析データは網羅的に得られることから非常に高次元になると考えられる。ランダムフォレストには、サンプル数が少なく説明変数が多い場合でも、過学習が発生しにくく、推定精度が低下しにくい特徴がある。この特徴によって、分析データの次元数に左右されにくい分析が行うことができる。

2　事前知識

2.1 和牛と品質評価

和牛とは日本在来の牛と国外の品種を配合して作られた品種群であり黒毛和種、褐毛和種、日本短角種、無角和種の 4 品種を指し、外国産の牛肉にはない筋肉に脂肪が混ざった霜降り肉になりやすいという特徴がある。外国産の牛肉との差別化を図るため、より脂肪交雑しやすくなるように品種改良が進められ高品質かつ高価格で生産、販売される傾向が強まっている。和牛の肥育農家はより良い品質を生み出す種を買い取り、肥育方法の経験を積むよう日々努力をしている。肥育農家は育てた子牛を食肉センターなどに出荷し、食肉センターで枝肉として加工され枝肉の成績情報と共に競売にかけられる。枝肉とは頭部、四肢、尾と腎臓以外の内臓を取り除いたあとの肉体のことである。和牛の品質の指標は数多く存在するが特に重要とされる枝肉重量、BMS(Beef Marbling Standard)、歩留基準値、バラの厚さ、皮下脂肪厚、ロース芯面積の 6 つの形質があり、主要 6 形質と呼ばれる。主要 6 形質の中でも枝肉重量は和牛の出荷総量を決めるため特に重要な指標とされている。肥育農家は高い肉質を維持しながらできるだけ枝肉重量が多くなるように肥育を行っている。

2.2 ランダムフォレスト

ランダムフォレスト[1]とは Breiman によって提案された決定木[2]を弱学習器とするアンサンブル学習を行う教師あり機械学習のアルゴリズムである。弱学習器とは単独で用いると精度の低い推定手法のことを指し、アンサンブル学習はこれらの弱学習器を多数組み合わせることで予測性能を向上させることができる。ランダムフォレストの場合、ランダムな復元抽出によって学習器のデータセットを作成する、ブートストラップサンプリングを用いて弱学習器である決定木を独立に多数作成する。ランダムフォレストにおける決定木の各ノードの分割基準は、どの説明変数を用いるか、選択した説明変数の連続した値をどこで 2 分するかの閾値の、2 つの要因によって決まる。末端ノードを作成するまで分割を繰り返し決定木は作成される。回帰を行う場合には各決定木に予測に用いるデータを与え、分割基準によって木を辿り、到達した葉ノードの分割基準で選択された変数が属するサンプルの目的変数の値が各決定木の回帰結果となる。この回帰結果の平均がランダムフォレストにおける回帰結果となる。

ブートストラップを用いて複数の決定木の平均を用いて予測を行うため、各決定木間の相関が下がり汎化性能が高まる。ランダムフォレストは決定木というモデルをアンサンブル学習させることで高い予測精度を実現している。ランダムフォレストの特徴として以下の 3 つが挙げられる。

*1 和歌山大学大学院システム工学研究科
*2 近畿大学大学院生物理工学研究科
*3 和歌山大学システム工学部

(a) サンプル数に対して説明変数が多くても過学習が発生しにくい
(b) ニューラルネットなどのディープラーニングに比べてサンプル数が少なくても一般に回帰性能が低下しにくい
(c) 説明変数の重要度を計算できる

特徴(a)と特徴(b)によってサンプル数が少なく説明変数が多いようなデータセットに対しても、回帰性能の低下度合いを低減しつつ高い推定精度を保つことができる。特徴(c)の重要度とは、目的変数に対して、各説明変数の影響の強さを相対的に表した値である。重要度は決定木の回帰にどれほど寄与したかの目安として考えられる。ランダムフォレストにおいては、重要度を計算した結果が 0 に近いほど推定に無関係な変数であると考えられる。したがって、重要度の低い変数を削除し再度ランダムフォレストを学習させることで推定精度が向上すると考えられる。このような重要度による説明変数の絞り込みはランダムフォレスト適用時に一般に広く用いられている。

3 提案手法

3.1 データ形式

提案手法では、説明変数として、導入時体重と、タンパク質発現量等の数多くの項目を含む分析データを想定する。導入時体重とは、繁殖農家から肥育農家に仔牛が販売された時点（導入時）での、その仔牛の体重である。また、本研究では、導入時から、例えば２～３ヶ月おきに何らかの牛のサンプルを取得し、そのタンパク質発現量や遺伝子発現量等の網羅的分析結果を得ることを想定しており、この分析データが説明変数となる。予測対象の形質としては、枝肉重量のみを想定する。

サンプルとなる和牛の集合をBとおく。和牛$b \in B$の導入時体重をW_bと書く。分析データに含まれるタンパク質等の項目の集合をPとおき、牛B_bにおける、項目$p \in P$の値を$v_{(b,p)}$と書く。なお、本研究で用いる分析データは、先述の通り経時的データであるが、提案手法の内部では取得された時期は考慮せず、例えば同じタンパク質でもサンプルの取得時期が異なれば異なるデータ項目として扱う。目的変数は枝肉重量であり、牛$b \in B$の枝肉重量をC_bと書く。

便宜的に、全ての牛$b \in B$に対する導入時体重の値の集合をW、枝肉重量の集合をCと書く。また、全ての牛$b \in B$、全ての項目$p \in P$に対する分析データの値$v_{(b,p)}$の集合をVと書く。

3.2 研究アイデア

本研究では、先述の通り、導入時体重と分析データを用いて、屠殺時の牛の枝肉重量を予測する手法を提案する。本応用例においては、あまり深い分析がなされておらず、まずは一般的な方法論を用いることが妥当である。例えば、重回帰分析や機械学習を用いた予測が考えられる。しかし、これらの方法では、それぞれの予測手法の性質が反映された予測となり、複数の手法の性質を利用することができない。アンサンブル学習のように複数の機械学習法を組み合わせた学習方法も存在するが、この方法では全ての予測結果の平均をとるため、個別の手法の特徴を十分に活かすことができない。

一方、和牛の枝肉重量に関しては、導入時体重と 0.7～0.8 を超える非常に大きい相関があることが知られている。この相関を活かした予測手法を構築できれば、従来手法よりも精度の高い予測が可能になると考えた。このためには、複数の手法の平均をとるのではなく、まず相関によって説明できる部分については単回帰により予測し、単回帰で予測できない誤差にあたる部分を、分析データに機械学習を適用することにより補助することで、より精度の高い予測が可能であると考えられる。

適用する機械学習法としては、ランダムフォレスト法を選んだ。説明変数となる分析データは網羅的に得られることから非常に高次元になると考えられる。ランダムフォレストには、サンプル数が少なく説明変数が多い場合でも、過学習が発生しにくく、推定精度が低下しにくい特徴がある。この特徴によって、分析データの次元数に左右されにくい分析が行うことができる。

以上のように、本研究では相関を活かした単回帰によって枝肉重量を予測し、発生する誤差をランダムフォレストで予測し、置き換えることで精度の高い予測を目指す。

3.3 回帰分析とランダムフォレストを組み合わせた枝肉重量予測手法

提案手法は、導入時体重と分析データから、枝肉重量を予測する。提案手法は学習フェーズと予測フェーズに分かれる。学習フェーズでは、回帰分析のモデルパラメータと、ランダムフォレストのモデルパラメータを決定する。学習フェーズは次の２段階の手順で構成される。

手順 1 導入時体重を用いて、単回帰分析により回帰分析のパラメータを推定する。
手順 2 単回帰分析の予測値と真値の差を目的変数として、分析データを入力し、ランダムフォレスト法のモデルパラメータを推定する。

手順 1 では、導入時体重Wと枝肉重量Cを用いて、導入時体重と枝肉重量の単回帰分析を行う。単回帰分析のモデル式を$C = \alpha W + \beta$とすると、手順 1 によってモデルパラメータαとβ が求まる。

手順 2 では、分析データVを説明変数、手順 1 で計算された予測値と真値の差を残差とし、残差Rを目的変数として、ランダムフォレスト法のモデル学習を行う。ここで牛bの残差$r_b \in R$は、手順 1 で計算されたモデルにおける牛bの枝肉重量の予測値$C'_b = \alpha W + \beta$を用いて、$r_b = C_b - C'_b$で表される。モデル学習により、ランダムフォレストの予測に用いる回帰木の集合が求められる。

一方、予測フェーズでは、学習フェーズで学習した各モデルに、予測に用いるサンプルを与えることで枝肉重量を予測する。推定フェーズは次の 3 段階の手順で構成される。

手順 1 学習フェーズの手順 1 で求めたモデル式に導入時体重を与え、単回帰分析による枝肉重量の予測値を求める。
手順 2 ランダムフォレスト学習器に分析データを入力し、残差予測値を求める。
手順 3 手順 1 で求めた枝肉重量の予測値に残差予測値を足した値を最終的な枝肉重の量予測値とする。

手順 1 では、予測に用いる牛サンプルb'の導入時体重の値$W_{b'}$を、学習フェーズで求めたモデルパラメータとモデル式を用いて、枝肉重量の予測値$C'_{b'}$が$C'_{b'} = \alpha W_{b'} + \beta$と求まる。

手順 2 では、学習フェーズで求めたランダムフォレストの各回帰木に牛サンプルb'の分析データ$V_{b'}$を与えることで残差予測値$r'_{b'}$が求まる。

手順 3 では、手順 1 と手順 2 で求めた$C'_{b'}$と$r'_{b'}$を加算することで提案手法による牛b'の枝肉重量予測値$C'_{b'} + r'_{b'}$を求める。

以上の学習フェーズと予測フェーズによって枝肉重量を予測する。

4 評価

4.1 入力データ

本研究で分析に用いたデータは以下の通りである。対象サンプルは同一農家で肥育された和牛 51 頭のデータである。入力となるデータは導入時体重、分析データとしてタンパク質発現量、枝肉重量である。タンパク質発現量とは検査対象中に含まれるタンパク質の量のことであり、一般的には質量や分子量など複数の対象が存在するが本研究では質量を対象とする。タンパク質発現量は牛の血清 1 μLに含まれるタンパク質の質量を SWATH-MS 法[3]と呼ばれる質量分析解析法を用いて網羅的に同定し、135 種類のタンパク質の含有量を測定した。また、血清は牛の導入時から出荷する 30 ヶ月齢まで 9、13、16.3、20.4、25、27 ヶ月齢の 6 回にわたって採取された。したがってタンパク質発現量データは 135×6 時期＝810 項目から欠損値を含む項目を取り除いた 702 項目となった。

4.2 評価方法

分析には Python のオープンソースの機械学習ライブラリである sklearn の sklearn.ensemble.RandomForestRegressor と sklearn.linear_model.LinearRegression を使用した。評価方法は提案手法である回帰分析とランダムフォレストを組み合わせた枝肉重量予測手法と既存の一般的な予測手法の精度を比較することで行う。評価指標には MAE (Mean Absolute Error)を用いる。精度の比較は、変数の数kを変化させ、それぞれのkで MAE を比較することで行う。

4.3 比較手法

比較手法は最も基本的な分析手法である重回帰分析、単純なランダムフォレストだけを行う手法の 2 種類である。ランダムフォレストを行う場合には、重要度による変数の絞り込みを変数がk個になるまで絞り込み再度ランダムフォレストを行った。重回帰分析を行う場合には、重回帰分析における変数の絞り込み手法である LASSO[4]を用いて変数の数がk個になるまで変数を絞り込み、推定を行った。重回帰分析の入力には目的変数として枝肉重量、説明変数として導入時体重とタンパク質発現量を用いた。また、各手法で MAE を計算する場合には 1 個抜き交差検証を行いサンプル数Nで平均をとった値を MAE とした。

4.4 評価結果と考察

表 1 は各手法における最低 MAE と、最低 MAE を記録した時の変数の数を示している。また、図 1 は 3 種類の各手法において、変数の数をk個に絞り込んで推定を行った際の MAE の推移である。図 1 は横軸が各手法に用いた変数の数(k)であり、縦軸は各手法によって推定された枝肉重量の MAE である。図 1 を見ると、同じ説明変数の数を用いていても提案手法は他の比較手法と比べて MAE が小さいことがわかる。これは同一条件（使用するデータセット、用いる変数の数）ならば提案手法が比較手法に対して推定精度において優れているということが言える。図 1 を見ると、提案手法とランダムフォレストでは、変数の数が 6 から 11 で最小の MAE を記録するまで徐々に MAE が減少している。これは重要度による説明変数の順位付けを行ったことにより、上位に MAE を減少させる能力を持つ説明変数を集める

表 1：最小 MAE 記録時の変数数

仕様手法	最小 MAE	変数数
提案手法	21.9617	6
ランダムフォレスト	25.6906	11
重回帰分析	24.3588	8

表 2：比較手法の単純なランダムフォレストにおける重要度

変数名	重要度
導入時体重	11.16
タンパク質 A	1.644
タンパク質 B	1.529
タンパク質 C	0.8072

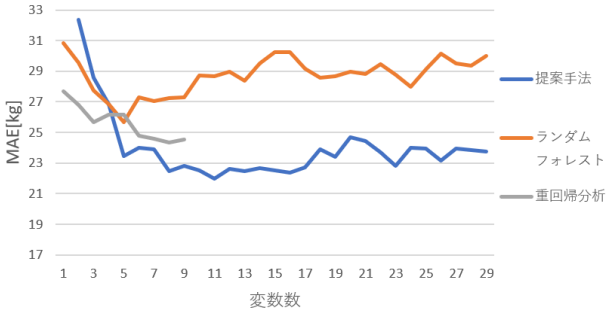

図 1：3 手法の MAE 推移

ことができたことを表す。つまり、重要度による変数の絞り込みは MAE の推定精度を向上させる役目を果たしている。また、説明変数を 700 項目以上使用しているにもかかわらず最小の MAE を変数の数が 5 から 10 で記録した後、MAE は停滞状態か右上がりの推移を示しており、これ以上使用する説明変数の数を増加させても MAE は減少していない。これは、多くある説明変数のうち、ほとんどが MAE を減少させる能力が低い変数であることを示している。

提案手法の性能がランダムフォレスト法よりも高い理由は、導入時体重を回帰を用いて扱ったことによる。導入時体重は枝肉重量と非常に高い相関があることが知られており、回帰分析に基づく予測は、回帰木に基づく予測とは異なる特徴を捉えている。実際に、ランダムフォレスト法により得られた説明変数の重要度を表 2 に示す。導入時体重の重要度が突出して高いが、図 1 を見ると、1 変数で学習した場合にはランダムフォレスト法の方が予測精度が高く、変数の数が増えると提案手法の予測精度が急激に上がる。これは、ランダムフォレスト法では導入時体重とタンパク質が共通の特徴要素を捉えているのに対して、提案手法では導入時体重を回帰分析で扱うことで、ランダムフォレスト法では捉えられない特徴を捉えたためと考えられる。

5 おわりに

本研究では、回帰分析とランダムフォレストを組み合わせた和牛の枝肉重量の予測手法を提案した。また、既存手法と枝肉重量の予測精度を比較することで提案手法の性能評価を行った。和牛の 6 つの主要な評価指標のうち、枝肉重量は導入時体重と相関が高いことが判明していた。そこで回帰分析により導入時体重の影響力を固定し、その他の要素をタンパク質の発現量によって説明しようと試みた。

評価では、提案手法に実データを適用することで提案手法の予測精度を調べた。重要度によって変数の順位付けを行い、順位の高い変数から提案手法に適用することにより、

高い予測精度を実現した。また、単純なランダムフォレストと重回帰分析を比較手法とし、提案手法との予測精度の比較を行った。その結果、提案手法が最も高い予測精度を示すことを明らかにした。

今後の課題としては、他のデータセットへの適用と、血清の採取時期による予測精度の評価の 2 点が挙げられる。今回の結果が本データセットに特有の性質ではなく、和牛全体に適用できる普遍的な性質であることを示すために、複数のデータセットでの検証が必要である。後者の採取時期の考慮は、提案手法の実用性を検証するために必要である。本予測結果を活用するためには、できるだけ早期に確度の高い予測が必要となる。つまり、今回の評価のように 6 時期のタンパク質発現量全てを用いるのではなく、より早期のデータのみに絞った場合の予測精度を調べることが重要である。

謝辞

本研究は日本中央競馬会畜産振興事業の支援を得た．ここに記して謝意を示す.

参考文献

[1] L Breiman:“RandomForst,” Machine Learning, vol.45, pp5-32(2001)

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone.: “Classification and Regression Trees," Chapman and Hall/CRC, Published January 1, pp.168 (1984)

[3] LC. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter,R. Bonner, and R. Aebersold: “Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a newconcept for consistent and accurate proteome analysis," Molecular & Cellular Proteomics (2012)

[4] R. Tibshirani:“Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, Series B (Methodological),Vol. 58, No. 1, pp. 267-288 (1996)