

RFE を用いた植物収穫時品質に関与する特徴量分析手法

中西豪太¹・峰野博史^{2,3}

¹ 静岡大学大学院総合科学技術研究科, 〒432-8011 静岡県浜松市中区城北 3-5-1

² 静岡大学大学院情報学領域, 〒432-8011 静岡県浜松市中区城北 3-5-1

³ 静岡大学グリーン科学技術研究所, 〒422-8529 静岡県静岡市駿河区大谷 8-3-6

Feature analysis method related to plant harvest quality using RFE

Gota Nakanishi¹ and Hiroshi Mineno^{2,3}

¹ Graduate School of Integrated Science and Technology, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011, Japan

² College of Informatics, Academic Institute, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011, Japan

³ Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, Shizuoka, 422-8529, Japan

Keywords: feature engineering, recursive feature elimination, plant harvest quality

DOI: 10.24480/bsj-review.11c4.00193

1. はじめに

農業従事者の高齢化に伴い、高水準な日本の農業技術が新規就農者に継承されずに喪失してしまうことが懸念されている。特に高糖度なトマトなどブランド価値の高い農作物を栽培する技術は熟練農家の長年の経験や勘によって培われており、会得するには長い年月を要する。そのため、高品質な農作物を栽培する技術は新規就農者に継承されることなく失われてしまう可能性が高い。この課題を解決するため、熟練農家が経験則と勘に基づいて判断していた多様な環境条件と植物の因果関係を明らかにする研究が行われている。例えば、光強度条件など栽培環境条件の違いが収穫時品質に与える影響の分析が行われている（浜本ら 2010, 東出 2018, 望月ら 1999）。ただし、実際の現場では、植物の生育状態や生育過程、時間帯といった考慮も行われており、そのようなデータを長期間かつ高品質に経時計測することは困難なため、栽培技術の形式知化を十分に実現できているとは言い難い。また、栽培中の環境条件は多岐に渡るため、様々な環境条件と収穫時品質を網羅的に検証することも困難である。そのため、新規就農者が熟練農家の持つ栽培技術を継承することは一筋縄ではいかない。

そこで、収穫時の品質に影響の大きい特徴量をデータドリブンで機械的に選択できないか試みた。生育状況と時間帯を重畳した経時特徴量データを算出（中西ら 2018, 水野ら 2018）し、決定木ベースの回帰手法を再帰的に用いて特徴量選択を適用することで、膨大な特徴量の中から収穫時の品質に影響の大きい特徴量を選択する。選択された特徴量を用いて収穫時の品質を推定できれば、生育状況と時間帯を考慮した分析の可能性を示すことができる。

2. 関連研究

2-1. 逐次変数選択法 (Stepwise 法) (Bendel et al. 1977)

変数群から重要な変数を選択する代表的な手法として、逐次変数選択法がある。逐次変数選択手法とは、変数増加法 (Forward stepwise selection) と変数減少法 (Backward stepwise selection) を組み合わせた変数選択手法である。変数増加法は最初に全ての変数に対して、モデルに加えた場合の p 値などのモデル自体の評価指標となる統計量を算出する。ここで、 p 値を変数選択の評価指標と仮定すると、事前に決定された閾値を満たす指標のうち、最小の値を持つ変数をモデルに加えてモデル構築をする。この処理を繰り返してモデルに含まれる変数を増加させていく。閾値を満たす指標が算出されなくなった時点で変数の追加を終了する。このときの変数の組み合わせを最良の組み合わせとする。一方、変数減少法は、変数増加法とは逆に、最初に全ての変数をモデルに取り込んだモデルを作成する。その後、モデルに含まれる変数のうち、決められた閾値を満たさない指標を持つ変数をモデルから除去していき、閾値を満たさない指標が算出されなくなった時点で変数の除去を終了する。逐次変数選択法は、最初に変数増加法と同様に変数の追加を行う。変数の追加のたびに、モデルに既に含まれている変数の中で変数減少法と同様に除去が可能な変数があれば除去する。その後、追加及び除去する変数がなくなるまで変数の選択を行い、最終的に残った変数を最良の変数の組み合わせとする。しかし、逐次選択法は、より少ない変数でより効率的に予測することが目的であり、目的変数との因果関係を考慮した変数選択を行わないため、選択された変数と目的変数の因果関係の分析には向いていない。また、変数の取捨選択を繰り返すため、重要な変数の最適な組み合わせを得ることも困難である。

2-2. 決定木ベースの機械学習手法を用いた変数選択

Random Forest (Breiman et al. 2001) や XGBoost (Chen et al. 2016) といった機械学習手法は、異なる決定木を多数作成し、その結果の平均値を求めることで、決定木の欠点である過剰適合を抑制するアンサンブル学習手法である。Random Forest は推定時に使用した特徴量の重要度を算出可能であり、重要度に基づいた特徴量選択が可能である。Random Forest の変数の重要度は、各決定木における変数の重要度の平均値によって算出される。まず、無作為にデータを選択して決定木を作成し、作成した決定木で使用されている 1 つの変数に関して、データの並び順をランダムに変更する。並び順の変更前後で、決定木の精度を比較し、大幅な精度の変化が観測された場合、重要な変数とする。最後に、多数の決定木にて同様にその結果の平均値を取り、Random Forest の変数の重要度とする。Random Forest の各変数の重要度は、変数全体から見た相対的な値であるため、変数の重要度を 0 にするスパースな推定を行わない。そのため、変数選択を行うには、閾値など人手を介した重要な変数を選別が必要となる。また、モデル構築時の重要度は、全特徴量に対する相対的な値であるため、一度の特徴量選択では重要な特徴量のみを選択することは困難である。これら課題を解決する手法として、モデル構築と特徴量削減を再帰的に行う RFE (Recursive Feature Elimination) (Guyon et al. 2002) という手法が提案されている。RFE は、モデル構築時の特徴量の重要度を基に特徴量を削減し続けることで、誤差指標が最良時の特徴量を選択することができる。

2-3. 正則化項を用いた変数選択 (廣瀬 2016)

遺伝子解析など、変数の次元数がデータセット数に比べて遥かに大きいという課題に対処できる統計手法として、正則化項によるスパース推定を用いた変数選択手法がある。正則化項を用いたスパースな推定を行う代表的な手法として、Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani et al. 1996) 回帰や Elastic Net 回帰 (Zou et al. 2005) などがある。Lasso 回帰モデルは、互いに相関の高い変数群が含まれている場合、推定時に相関の高い変数の中の 1 つだけが選択され、他の変数は回帰係数が 0 と推定され、選択されないという課題がある。つまり、相関の高い変数群の中で選択される変数は、推定を行うごとに変化してしまい、Lasso 回帰による変数選択は不安定とされている。この Lasso の課題を解決するため考案された回帰手法が、Elastic Net 回帰である。Elastic Net 回帰は、Ridge 回帰 (Hoerl et al. 1970) と Lasso 回帰を混合した回帰手法である。Ridge 回帰は、相関の高い変数群を考慮できるため、Lasso 回帰の正則化項と Ridge 回帰の正則化項の強さのバランスをとることで、Lasso 回帰の課題を解決する。また、Lasso 回帰と Elastic Net 回帰においては、自動で変数の選別を行うため、閾値の決定など人手の介入なくとも変数選択を実現できる。しかし、Lasso 回帰や Elastic Net 回帰は線形的な回帰手法であるため、対象が植物の成長といった複雑な変数との非線形性な関係性を持つ場合には不向きな場合が多い。

2-4. 関連研究のまとめ

以上のように、変数選択においては、 p 値などの統計的な指標を算出した変数に対し、事前に設定した閾値を用いて選別することで変数選択を行う手法が多い。しかし、重要な変数を余すことなく、選別することが可能な閾値を決定することは難しい。また、Lasso 回帰や Elastic Net 回帰では非線形な関係性の考慮が困難であるため、非線形な関係性を考慮可能な Random Forest や XGBoost といった機械学習手法によって誤差が最小となるよう再帰的にモデルを構築し、特徴量を削減する手法である RFE を適用することで、重要度の高い特徴量を機械的に選択できる可能性が高い。

3. RFE を用いた植物収穫時品質に関与する特徴量分析手法

3-1. 概要

植物の生育状態と時間帯を考慮して収穫時の品質に影響の大きい経時的な特徴量を機械的に選択する手法を検討した。特に、目的変数に対する特徴量の重要度が算出可能である決定木ベースの回帰手法を再帰的に用いる特徴量選択手法を植物栽培環境のような対象に適用できるか検証した。代表的な決定木ベースのアンサンブル回帰手法として、Random Forest や XGBoostなどを想定する。決定木ベースのアンサンブル回帰手法では、推定時に使用した特徴量の重要度を算出可能であるため、生育状態や栽培した時間帯を考慮した経時的な特徴量に対して、重要度を指標とした特徴量選択を行うことで、植物栽培環境における収穫時の品質に影響の大きな特徴量を機械的に選択する特徴量分析手法を検証する。

本手法の概要を図 1 (a) に示す。本手法は図 1 (a) に示すように、大きく A：特徴量算出と B：特徴量選択から構成される。A：特徴量算出では、経時的な特徴量データを栽培対象の農作物の生育ステージ毎に分割した後、時間帯ごとに分割する。次に図 1 (b) に示すように、分割された各経時特徴量データに対して基本統計量を算出することで、植物の生育状態と栽培した時間帯を考慮した特徴量の算出を行う。また、栽培期間を通じた特徴量として、温度や日射量の積算値や、栽培開始からの経過日数など、栽培した季節の考慮が可能な特徴量を算出する。B：特徴量選択では、A：特徴量算出において算出された多数の特徴量に対し、機械学習を用いて再帰的に特徴量を削除する手法である RFE を用いて、収穫時の品質に影響の大きい特徴量の選択を行う。

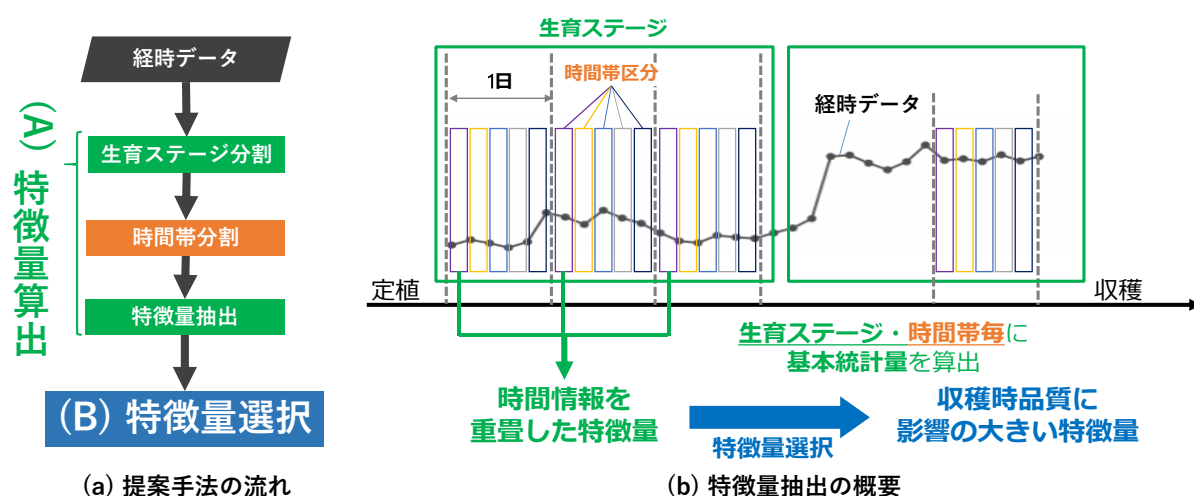


図 1 特徴量選択手法の概要

Fig. 1 Outline of the feature analysis

3-2. 特徴量算出

特徴量算出では、経時特徴量データと生育状態、栽培した時間帯情報を重畳するため、経時特徴量データを生育状態と栽培時間帯ごとに分割する。分割の際には植物の状態や栽培した日のイベントを指標とする。生育状態は植物の開花から収穫までの栽培期間を対象とし、果実の生育ステージを指標に用いて分割する。栽培期間中の植物のイベントである開花や収穫を基準に分割を行うことで、果実の個体差を考慮した生育状態の分割が可能であると考ええる。また、栽培時間帯に関しては、日の出、日の入りを分割の指標に用いる。一般に植物は光合成によって果実の成長が促進される。そのため、日の出と日の入りを指標に用いることで、光合成の不可欠な要因である日射量の影響が考慮可能となると考える。次に生育状態と栽培時間帯に分割した各経時特徴量データに対して基本統計量の算出を行う。一般に熟練農家は特定の 1 日の植物の変化ではなく、長期的な変化を考慮した栽培技術を用いるため、基本統計量には大域的な変化を表現可能な積算、平均、最大、最小を用いることとした。また、栽培を通じた特徴量として、植物の成長の指標に用いられる栽培期間の積算温度と、積算日射量や定植日からの日数を算出し利用することとした。

3-3. 特徴量選択

Random Forest や XGBoost などのモデル構築時の特徴量の重み付けが可能な機械学習手法を用いてモデルを構築する。構築時に重要度の低い特徴量を削減する処理は、誤差指標を基準として再帰的に繰り返すことで特徴量選択を行う手法である RFE を特徴量算出で算出された特徴量に対して適用することで、機械的に目的変数に設定した収穫時の品質に対して重要な特徴量の選択が可能であると考えられる。また、特徴量選択の信頼性を高めるため、RFE で使用する外部学習器を複数使用することで、学習器別での特徴量選択結果の違いを考慮する。

4. 基礎評価

実際に栽培した際のデータを用いて生育状態と栽培時間帯を重畳した特徴量の算出と、算出した特徴量から重要度の高い特徴量の選択、選択された特徴量を用いた収穫時品質の推定を実施した。本基礎評価では、2016 年 7 月から 2017 年 12 月にかけて複数のトマト (CF 桃太郎ヨーク) 栽培農場にて低段密植栽培された 54 作分の栽培データを使用した。また、栽培期間に収集されたセンサデータとして、温度 ($^{\circ}\text{C}$)、飽差 (g/m^3)、日射量 ($\text{J}/\text{m}^2/\text{s}$)、 CO_2 濃度 (ppm) を 1 分周期で測定した環境データを使用し、収穫時の品質の 1 つである収量は選果機を用いて 1 日毎に計測された値を使用した。特徴量算出での生育ステージの区分は、農学者の知見をもとに定植日から収穫開始日までを 5 等分し、生育ステージ I から生育ステージ V と設定した。栽培した時間帯の区分は、栽培した農場の日の出時刻と日の入り時刻を基準に、日の出時刻から日の入り時刻を 3 等分し、それぞれを時間帯 1 から時間帯 3 とした。また、日の入り時刻から翌日の日の出時刻までを 2 等分し、それぞれを時間帯 4、時間帯 5 と区分した (表 1)。また、特徴量を算出する際の基本統計量には、大域的な変化を表現可能である積算、平均、最大、最小を用いた。以後、図中の温度、飽差、日射量、 CO_2 濃度は、それぞれ temp, vpd, solar, CO_2 と表記し、積算、平均、最大、最小はそれぞれ sum, ave, max, min と表記する。

特徴量選択で用いる RFE の外部学習器には、モデル構築時に特徴量の重み付けが可能な学習器である必要があるため、Random Forest, XGBoost, LightGBM (Wang et al. 2017) を用いて特徴量選択結果を比較した。選択された特徴量を用いた予測では、収穫開始日から 4 週間後までの 1 週間毎の総収量を予測対象とした。予測時に使用する学習器は

Random Forest, SVR (Support Vector Regression) (Drucker et al. 1997), XGBoost を用いた。また、予測結果の評価指標として、平均絶対誤差 (MAE : Mean Absolute Error) (式 1) と平均二乗誤差平方根 (RMSE : Root Mean Squared Error) (式 2), 決定係数 (coefficient of determination) (式 3) を用いた。実装は、Python3.7, scikit-learn

表 1 生育ステージと時間帯区分方法

Table 1 Growth stage and time zone division

生育 ステージ	想定する 生育状態	時間 帯	範囲
I	着花	1	日の出から日 没 (3 等分)
II	開花前	2	
III	開花	3	
IV	着果	4	日没から日の 出 (2 等分)
V	果実肥大	5	

(Pedregosa et al. 2011) 0.21.0 を使用し、パラメータチューニングには Optuna を使用した。

$$MAE = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3)$$

4-1. 相関分析

特徴量算出にて算出された特徴量の有効性を検証するため、算出された特徴量 400 次元 (センサ値 (温度, 飽差, 日射量, CO₂ 濃度) × 生育ステージ (I ~ V) × 栽培した時間帯 (1~5) × 基本統計量 (積算, 平均, 最大, 最小)) と収量との相関行列を図 2 に示す。

栽培期間の生育ステージⅢから生育ステージⅤの時間帯 1, 時間帯 2 の日射量, 温度との正の相関が特に高いことが分かる。生育ステージⅢは本実験では, 開花時期を想定した生育ステージである。そのため, 正の相関が高いことから, トマトは開花後の光合成の促進が収量増加に効果的である (東出 2010) という農学的な知見と一致することが分かる。また, 時間帯に着目すると, 時間帯 1, 時間帯 2 の日射量, 温度との相関が高いことが分かる。一般に光合成速度は午前中に最大値に達し, その後徐々に低下するという推移を示す。時間帯 1 は日の出時刻を指標として区分された時間帯である。そのため, 日の出により光強度が増加したことで, 光合成速度が光飽和点付近まで上昇したことで, 光合成の促進に繋がったと考えられる。また, 光合成には, 光の他に温度, CO₂ 濃度が重要である。これらの要因は光合成を行う上でそれぞれ限定要因となるため, 時間帯 1, 時間帯 2 において, 温度との相関が高いという結果については, 光と温度が光合成を行うために十分供給されたことで, 光合成速度が大きくなり, 光合成が促進され収量の増加に繋がったと考えられる。

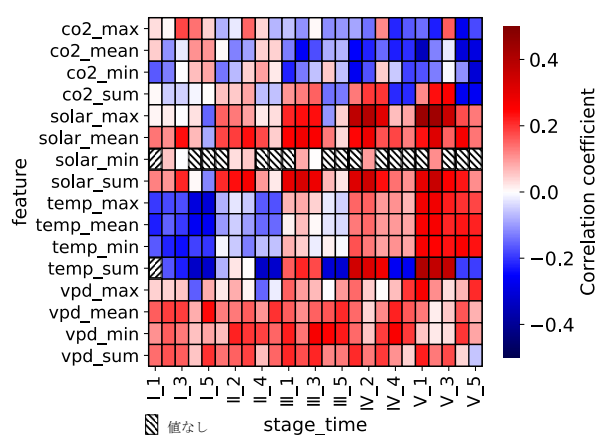


図 2 算出された特徴量と収量との相関

Fig. 2 Correlation between calculated feature quantity and yield

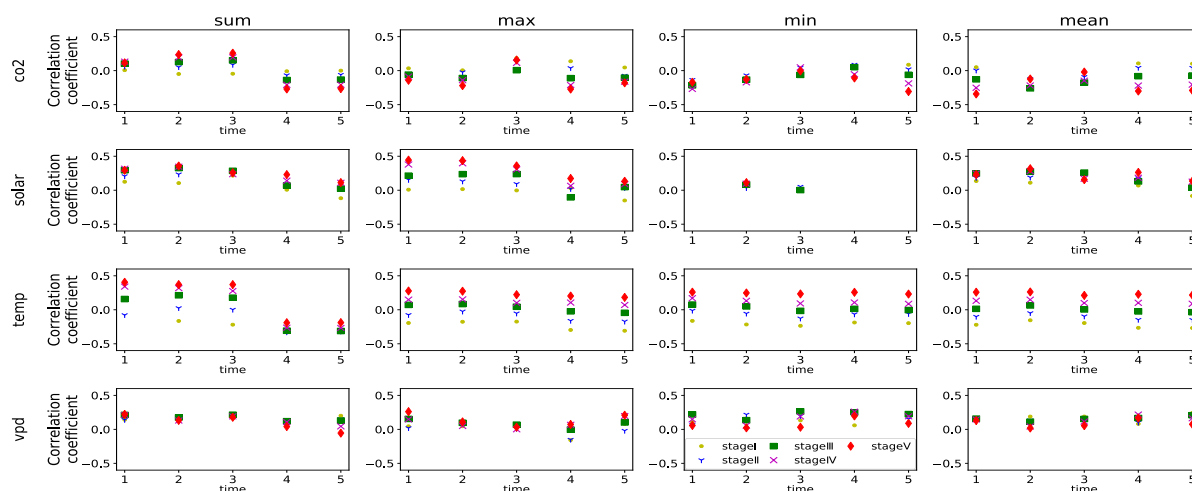


図 3 生育ステージ毎の相関

Fig. 3 Correlation by growth stage

算出した特徴量と収量との相関を、生育ステージ毎に図 3 に示す。図 3 における 2 行目に示す日射量に着目すると、栽培期間を通して時間帯 1 から時間帯 5 の積算値、最大値との相関が高い。特に開花前の積算日射量は収量との相関が高い(Pedregosa et al. 2011) ことが報告されており、開花時期を想定している生育ステージⅢの日射量の積算値との相関が大きい結果は妥当であると考ええる。また、前述の通り、果実の肥大が進む生育ステージⅣ以降の日射量との相関が高い点からも、算出された特徴量は時間情報の重量が実現できていると考える。一方、図 3 の 4 行目に示す飽差に着目すると、栽培期間を通して飽差の最小値との相関が大きい事がわかる。つまり、飽差が小さいことで葉の気孔が開いた際の蒸発散速度が上昇し、収量に好影響だったと考えられる。よって、算出された特徴量は植物栽培における生育ステージや一日の時間帯といった時間情報を考慮した特徴量と言え、栽培期間中の環境要因が収穫時品質に与える影響の分析に有効である可能性が示された。

表 2 RFE のパラメータ

Table 2 Parameters of RFE

パラメータ	値
step	1
min features to select	10
cross validation	5
scoring	mae
regressor	XGBoost, LightGBM, Random Forest

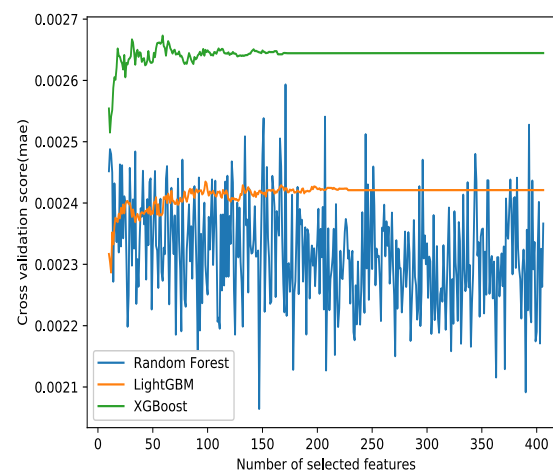


図 4 各学習器の特徴量選択推移

Fig. 4 Feature selection transition of each learner

4-2. 特徴量選択の検証

算出された特徴量から、収穫時の品質に影響の大きい特徴量を選択するため、RFEを用いた特徴量選択を行った。RFEにて使用する外部学習機にはXGBoost, LightGBM, Random Forestの3種類を用いた。RFEの評価には、収量を栽培した株数で除算し、1株あたりの収量とすることで、生産者の栽培規模の違いを考慮した目的変数とした。RFEの各パラメータを表2に、各学習器の特徴量選択推移を図4に示す。また、各学習器を用いた特徴量選択結果を図5から図7に示す。

図4より、Random Forestの特徴量選択推移が一定の範囲内で発散していることがわかる。Random Forestは無作為に選択された特徴量を用いて決定木を複数作成するアンサンブル学習であるため、特徴量の削減を繰り返すことで、決定木作成時の特徴量の組み合わせパターンが限定され、値が一定の範囲内で発散したと考えられる。図5に示すXGBoostによる特徴量選択結果から、生育ステージVの時間帯1と時間帯2の日射量が、重要度の高い特徴量として選択されたことが分かる。時間帯1、時間帯2は日の出時刻からの時間帯であるため、生育ステージVにおいて日の出による光強度の向上により、光合成速度が大きくなり、光合成が促進されたことが収量増加に繋がったと考えられる。また、同様に生育ステージVにおける日の出後の温度の積算値の重要度が高いという結果となった。植物の光合成促進を決定する要因である、温度、光強度、CO₂濃度のそれぞれが限定要因となるため、温度と光強度の両方が高まったことで収量が増加したと考えられる。植物は一般的に光合成速度が正午頃、つまり時間帯2に最大となり、その後は徐々に低下するという推移を示すこと（藤澤ら 2010）から、生育ステージVの日の出後から昼過ぎの光合成速度の増加が収量に対して重要であるといえる。生育ステージVは、

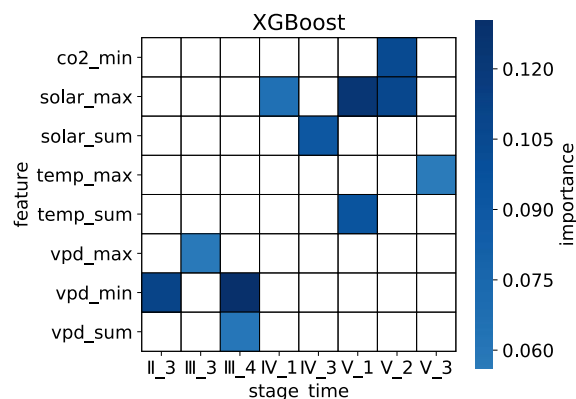


図 5 XGBoost による特徴量選択結果

Fig. 5 Feature selection result by XGBoost

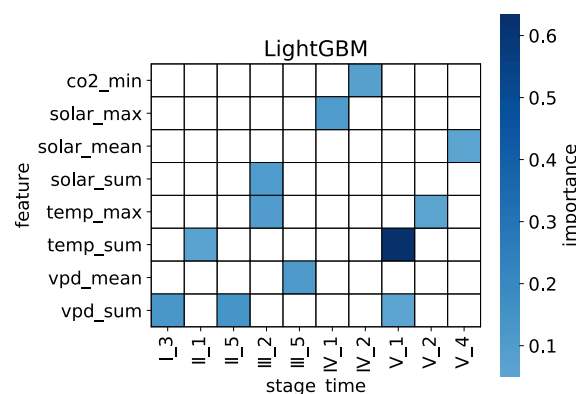


図 6 LightGBM による特徴量選択結果

Fig.6 Feature selection result by LightGBM

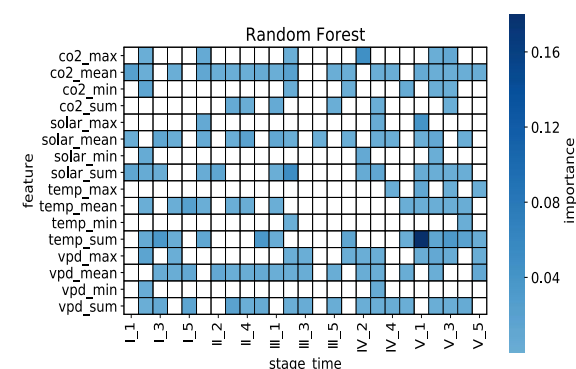


図 7 Random Forest による特徴量選択結果

Fig. 7 Feature selection result by Random Forest

果実が赤くなる完熟期を想定した生育ステージであるため、トマトは緑熟期に収穫時の7割程度の糖度を含有し、緑熟期以降に更に糖度が上昇する（石井ら 1994）。そのため、緑熟期として想定した生育ステージⅣ以降の光量を増加させることが、収穫時糖度を向上する要因になると期待できる。

一方、図6に示すLightGBMを用いた特徴量選択結果では、生育ステージⅤの時間帯1の温度の積算が重要な特徴量として選択された。この結果は、XGBoostによる特徴量選択結果と一致するが、日射量の最大値は選択されなかった。また、LightGBMによる特徴量選択結果では、生育ステージⅢの午前中の日射量の積算、温度の最大という光合成を促進させる要因が選択され、XGBoostによる特徴量選択結果と異なり、生育ステージⅢの環境要因を重要視していることがわかる。

図7に示すRandom Forestを用いた特徴量選択結果では、147次元の特徴量が選択され、他の学習器と異なる選択結果が散見された。このような学習器の違いによる特徴量選択結果の違いを考慮するためには、複数の学習器の結果を用いた特徴量選択が有効であると考えられる。

そのため、学習器毎の特徴量選択結果の違いを考慮できるよう、異なる学習器を用いたRFEの結果を組み合わせた特徴量選択を試みた。図9にXGBoostによる選択結果の重要度と、LightGBMによる選択結果の重要度をそれぞれ正規化し、平均値化した結果を示す。複数の学習器を用いた特徴量選択によって、単一の学習器による特徴量選択時の欠点である結果の信頼性の向上が期待できる。

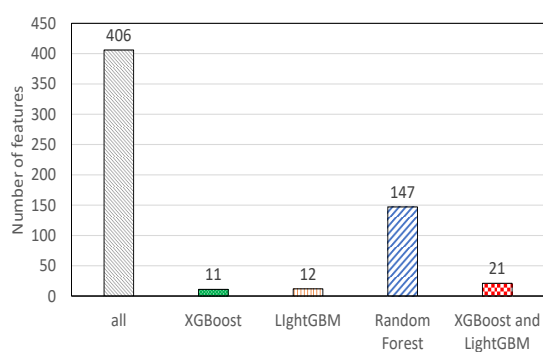


図8 学習器別の選択された特徴量数

Fig. 8 Number of selected features by learner

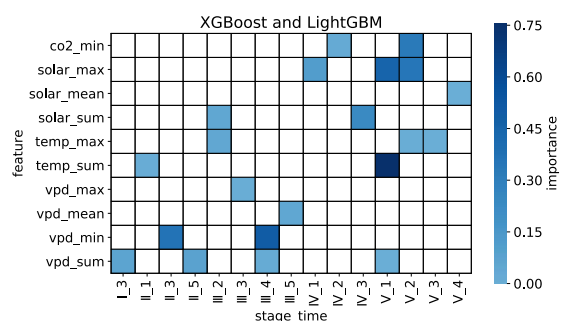


図9 XGBoost と LightGBM による特徴量選択結果

Fig. 9 Feature selection result by XGBoost and LightGBM

4-3. 選択された特徴量による収量予測

特徴量選択結果の検証として、選択された特徴量を用いた収量の予測性能を評価した。図 5, 図 6, 図 9 で示した各特徴量を用いた予測時の誤差と、全特徴を用いた推定誤差を比較することで、特徴量選択結果の妥当性を検証する。予測する収穫時品質として収量を設定し、栽培者やその栽培規模の違いを正規化するため、収穫開始日から 4 週間後までの 1 週間後ごとの収量 (kg) を株数で割った値を予測対象とした。

データセットに関しては、収集した 54 件の栽培データを、一般的に収穫が終了する収穫開始から 4 週間後まで 1 週間ごとに抽出し、それぞれ予測対象週に応じて生育状態と時間帯分割を行うことでデータの拡張を行った。拡張したデータに対して、予測対象週の偏りが発生しないよう、予測対象週ごとに学習データとテストデータを 8 対 2 の割合で分割し、データセットとした。収穫開始から 3 週間以内に収穫が終了した作に関しては、収穫終了週までのデータを用いた。

評価諸元と学習器のパラメータを表 3 と表 4 に、図 10, 図 11 に特徴量、学習器別の予測誤差を示す。全特徴量を用いた予測時誤差と比較して、RFE を用いて選択した特徴量のみを用いた場合でも同程度の誤差の予測が行えていることが分かる。また、RFE に使用した学習器に着目すると、単一の学習器による特徴量選択結果を用いた RFE (XGBoost) と RFE (LightGBM) では全特徴量を用いた予測と比較して予測精度として、MAE がそれぞれ約 12.4%, 約 10.1%, RMSE がそれぞれ約 10.5%, 約 8.0%悪化した。精度が悪化した原因としては、図 8 より選択された特徴量数がそれぞれ 11 次元, 12 次元と少

表 3 評価諸元

Table 3 Evaluation specifications	
項目	内容, 値
目的変数	収量(kg) / 株数(本)
number of cross validations	5
scoring index	RMSE
regressor	LightGBM, Random Forest, SVR, XGBoost

表 4 パラメータ探索範囲

Table 4 Parameter search range	
LightGBM	
min data in leaf	100 - 500
number of leaves	10 - 300
max depth	1 - 10
number of estimators	10 - 300
learning rate	0.01 - 0.2
Random Forest	
max depth	100 - 500
number of estimators	100 - 500
SVR	
penalty parameter C	2e-10 - 2e11
kernel	rbf
gamma	2e-10 - 2e11
epsilon	2e-10 - 2e11
XGBoost	
min samples in leaf	100 - 500
number of leaves	10 - 300
max depth	1 - 10
number of estimators	10 - 300
min samples to split	100 - 500

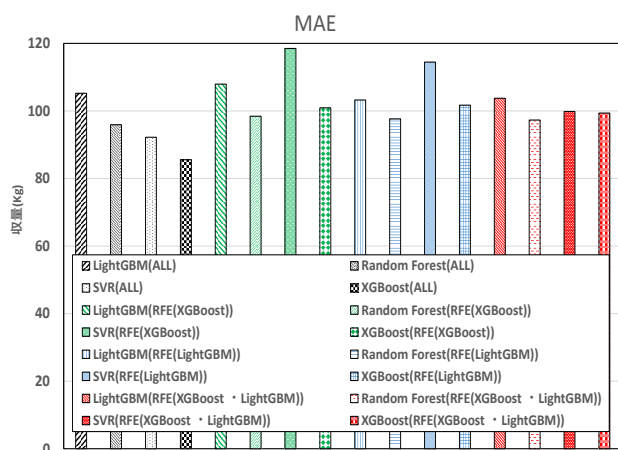


図 10 予測誤差 (MAE)

Fig. 10 Prediction error (MAE)

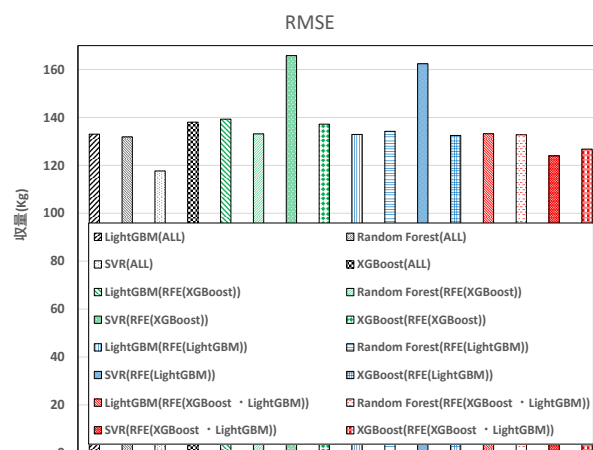


図 11 予測誤差 (RMSE)

Fig. 11 Prediction error (RMSE)

なく、特徴量選択時に過学習してしまったと考えられる。しかし、複数の学習器を用いた RFE による特徴量選択では、もとの特徴量から 385 次元の特徴量が失われているにもかかわらず、各学習器を平均して MAE が約 5.6%悪化, RMSE が約 0.7%向上と、単一の学習器による RFE と比較して性能を維持できていることが分かる。そのため、複数の学習器を用いた RFE によって選択された 21 次元による単純なモデルでの効率的な予測が行えたと考える。つまり、選択された特徴量は植物収穫時品質への影響の大きい特徴量であると言える。また、一般に熟練農家は植物の生育状態やその時間帯を自身の経験則から総合的に判断して栽培や環境制御を行うため、栽培期間中の特に重要度の高い要因を明らかにできれば、新規就農者の高品質な農作物栽培に役立つものと考ええる。

5. おわりに

本稿では、植物の栽培期間中の環境要因が収穫時品質与える影響や重要度を栽培データから明らかにするため、植物が成長とともに環境要因やその栽培時間帯が与える影響が異なることに着目し、植物の生育ステージや時間帯を考慮した特徴量を算出し、RFE による機械学習モデルベースの収穫時品質に影響の大きい特徴量選択手法を検討した。本手法を用いて実際に低段密植栽培トマトを栽培した際のデータを分析した結果、特徴量選択では午前中の日射量や温度といった光合成を促進させる要因が選択され、農学的な知見と一致する結果を得ることができた。また、選択された特徴量のみを用いて収穫時の品質である収量を予測した結果、算出された全ての特徴量を用いて予測した際と同程度の精度の結果を得られたことから、選択された特徴量は収量予測時に重要度の高い特徴量であり、この特徴量を考慮した予測や分析、栽培を意識することで新規就農者の栽培支援に役立てられると考える。

今後は、データセットの拡充だけでなく、灌水や降雨といった植物に対して影響の高い外部要因を含めた分析を行うことで、収穫時の品質予測の精度向上を目指す。特に栽培期間中の植物の生育状態を記録することで、より正確に生育状態を考慮した特徴量を検討できると考える。また、他の圃場にも本手法を適用することで、汎用性の評価も行うとともに目的変

数を糖度など他の収穫時の品質でも有効であるか検証を進めていく。

謝辞

本研究は JST さきがけ (JPMJPR15O5) の支援を受け実施されたものである。また、本研究にあたり、テラスマイル株式会社の金田千広様には農学的な観点からのアドバイスを頂いた。ここに感謝の意を示す。

引用文献

- Bendel, R.B., & Afifi, A.A. 1977. Comparison of stopping rules in forward “stepwise” regression, *Journal of the American Statistical Association*, Vol.72, No.352, pp46-53.
- Breiman, L. 2001. Random Forests, *Machine learning*, Vol.45, No.1, pp5-32.
- Chen, T., & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System, *KDD '16*, pp.785-794.
- Define-by-run hyperparameter optimization framework. <https://optuna.org/> (アクセス日 2019-06-26).
- Drucker, H., Burges, C.J.C., Kaufman, L., et al. 1997. Support vector regression machines, *Advances in Neural Information Processing Systems*, Vol.9, pp155-161.
- 藤澤弘幸ほか 2010. JM1, JM7, JM8 および M.9 台木がリンゴの葉の光合成速度に及ぼす影響, *園芸学研究*, Vol.9, No.2, pp.171-176.
- Guyon, I., et al. 2002. Gene selection for Cancer classification using support vector machines, *Machine learning*, Vol.46(1-3), pp.389-422.
- 浜本浩, 星 岳彦, 山崎敬亮ほか 2010. 3 段取りトマト栽培における群落内補光の時間帯が収量に及ぼす効果と補光の経済性, *植物環境工学*, Vol.22, No.2, pp.95-99.
- 東出忠桐 2010. 開花前の日射に基づいた夏秋トマトにおける週間収量変化の予測, *園芸学研究 別冊*, Vol.9, No.1, p13.
- 東出忠桐 2018. 施設トマトの収量増加を目的とした受光と物質生産の関係の利用, *園芸学研究*, Vol.17, No.2, pp.133-146.
- Hoerl, A.E., & Kennard, R.W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol.12, No.1, pp55-67.
- 廣瀬慧 2016. スパースモデリングとモデル選択, *電気情報通信学会誌*, Vol.99, No.5, pp392-399 (2016).
- 石井孝典, 藤野雅丈, 矢ノ口幸夫 ほか 1994. トマト品種の果実成分と熟度の関係, *東北農業研究*, Vol.47, No.1, pp.275-276(1994).
- 水野涼介, 柴田 瞬, 峰野博史 2018. 植物収穫時品質や収量に関連する経時特徴量分析手法の検討, 第 80 回全国大会講演論文集 2018.3, pp.121-122.
- 望月龍也, 石内伝治, 伊藤喜三男 1999. トマト果実における糖含量およびその栽培・環境条

- 件に対する安定性の品種間差異, 園芸学会雑誌, Vol.68, No.5, pp.1000-1006.
- 中西豪太, 水野涼介, 今原淳吾, 前島慎一郎 ほか 2018. 植物収穫時品質に關与する経時特
微量の検討, 情報処理学会マルチメディア, 分散, 協調とモバイル(DICOMO2018)シンポ
ジウム, pp.75-81.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, Vol.12, pp.2825-2830.
- Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical
Society: Series B (Methodological), Vol. 58, No.1, pp. 267–88.
- Wang, D., Zhang, Y., & Zhao, Y. 2017. LightGBM, Proceedings of the 2017 International Conference
on Computational Biology and Bioinformatics - ICCBB 2017, ACM Press, pp. 7–11.
- Zou, H., & Trevor H. 2005. Regularization and variable selection via the elastic net, Journal of the
Royal Statistical Society: Series B (Statistical Methodology), Vol.67, No.2, pp. 301–320.