O-001

大豆栽培環境の交互作用を考慮した GA を用いたハイブリッド特徴選択
Hybrid feature selection using GA in light of soybean growing environment interaction

大橋真琴[1)]　　大川剛直[1)]
Makoto Ohashi　Takenao Ohkawa

1　はじめに

近年，機械学習によりデータを解析し収量を予測することで, 効率的に栽培管理を行うスマート農業が進められている．予め収量がわかることで, 出荷の調整や労働力の確保が可能になる．こうして収量の予測結果を実際の圃場で活用するにむけて, 精度の良いモデルの開発が求められている．機械学習の精度改善にはしばしば特徴選択が用いられる．特徴選択では, データセットから有用な特徴を選択するため，解釈しやすいという利点があるが，情報損失が必ず生じる. また, 作物の生育は複雑であり, 栽培環境要因が複雑に絡み合い収量に影響を与える．よって, 単一の要素に由来する影響だけでなく, 複数の因子が組み合わさることで初めて現れる相乗効果である交互作用による影響も考慮すべきである．

そこで本研究では, GA を用いた特徴選択により交互作用に関する情報損失を抑えながら栽培環境と収量の関係性を抽出するとともに, 得られる収量予測モデルの精度改善を目指す．また, 収量の予測には, 栽培環境データの時系列性を考慮するため, Temporal Random Forests(TRF)[1] を利用する．さらに, GA による特徴選択と TRF による予測における計算時間が大きくなってしまうといった欠点に対し，χ 二乗フィルターによる特徴選択を併用することで対処する．図 1 に提案モデルの概要を示す.

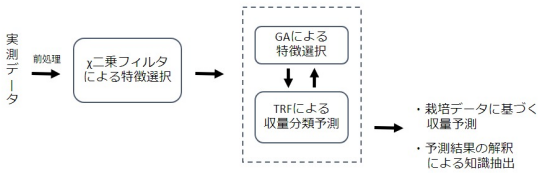

図 1　提案モデルの概要

2　収量分類手法

多収, または低収となる原因となった栽培環境を抽出するため, 気象条件に基づく収量分類問題を考える．また, 本研究で扱う気象データは多変量時系列データである．そこで, 分類器として特徴間の時間的位置関係を考慮する Temporal Random Forests(TRF) を用いる．これによりデータの時系列性を利用した分類が可能であるが, モデルが複雑であるため, 学習に時間がかかるといった問題が発生する.

1)　神戸大学大学院 システム情報学研究科
Graduate School of System Informatics, Kobe University

3　ハイブリッド特徴選択

特徴選択手法はフィルター手法, 埋め込み手法, ラッパー手法と大きく三つに分けられ, それぞれに利点と欠点が存在する. 本研究では, ①分類精度の改善②計算時間の短縮③交互作用を考慮したモデルを実現するため，χ 二乗検定と遺伝的アルゴリズム (GA) を用い, それぞれの欠点を補いあうハイブリッド特徴選択を行う.

3.1　χ 二乗検定

χ 二乗検定では, 説明変数の各特徴量と目的変数の独立性について検定を行う．評価する特徴量を t, 目的変数を c, データ数を N とすると, χ 二乗検定は以下の式で表される.

$$\chi^2(t, c_i) = \frac{N[P(t, c_i)P(\bar{t}, \bar{c_i}) - P(t, \bar{c_i})P(\bar{t}, c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c_i})}$$

χ 二乗の値が大きいほど, 観測値と期待値の差があるということを表しており, 目的変数への影響が大きい特徴量として選択される．この χ 二乗フィルター手法によって, 説明変数を目的変数への影響が大きいものに絞ることができ, ①分類精度の改善②計算時間の短縮が実現する．しかし, フィルター手法では特徴量を 1 つずつ評価するため, ③交互作用が考慮されない.

3.2　GA

GA では, 生物の遺伝の仕組みに倣い, 交差継承と突然変異を繰り返すことで最適解を探索する．特徴量選択において GA を用いる場合は, 染色体で各特徴量を用いる (1) か用いない (0) かを表現し, 適合度をモデルの予測精度とすることで, 適合度が大きくなる, すなわち予測精度が高くなるよう特徴量選択を行うことができる [2]．GA による特徴選択の手順を以下に示す.

- **STEP1**: 初代遺伝子をランダムに生成.
- **STEP2**: 遺伝子が示す特徴量サブセットに対し,TRF の分類精度を確認.
- **STEP3**: 分類精度を基に交叉・突然変異により次の世代の遺伝子を作成.
- STEP2.3 を繰り返す

この GA ラッパー手法によって, 最も分類精度が高くなるよう特徴量のサブセット単位で探索することができ, ①分類精度の改善③交互作用を考慮したモデルが実現する．しかし, GA の探索の過程で何度も機械学習を行うため, ②計算時間についてはかえって大きくなってしまう.

4　実験

大豆圃場それぞれにおける大豆の播種日から収穫日までの気象データ, 最終的な収量データなどの実データを用いて提案手法の評価を行う．本研究では, 農林水産省の収益力向上のための研究開発プロジェクトより提

供された全国の大豆圃場に関するデータから, 圃場ごとの位置情報, 播種日, 開花日, 収穫日, および収穫量を取得する．また, 農研機構メッシュ農業気象データシステムが提供する 1km × 1km のメッシュ気象データ [3] を利用することにより, 圃場の位置における一日ごとの気象データを取得する．全国の大豆圃場に関するデータは 2015 年から 2018 年の間に取得されたデータである．

4.1 データの前処理

提供されるメッシュ気象データから, 特徴間の相関を基に平均気温・降水量・風量・気温差を利用する．また, これらの気象データに対し, 機械学習の精度向上並びに実際の栽培管理のしやすさの観点から 2 種類の加工を行う．1 つめのステージ分割では, 時系列データである気象データの取得基準を大豆の生育ステージ毎とし, 本研究では 13 ステージと定めた．2 つ目の定性値化では, 気象データを平年値と比較しカテゴリ表現にすることで, 多変量データの粒度をそろえる [4].

4.2 提案モデル分類精度

表 1 各パートの組み合わせによる実験結果

	精度	実行時間	交互作用の考慮
TRF	0.827	8h	×
TRF × χ 二乗	0.83	3h	×
TRF × GA	0.842	200h	○
TRF × χ 二乗 × GA	0.842	25h	○

本研究では, ①分類精度②計算時間③交互作用の考慮の 3 点から手法の評価を行う．分類精度の評価指標としては Accuracy を用いる．また, 予備実験から, χ 二乗検定により選択する特徴量を全体の 50%とし, GA の 1 世代あたりの遺伝子数を 25・世代数を 50 と定めた．本稿で紹介した各パートの組み合わせによる実験結果を表 1 に示す.

表 2 様々な組み合わせのハイブリッド特徴選択による実験結果

	精度	実行時間	交互作用の考慮
RF	0.788	1h	×
RF × χ 二乗 × GA	0.796	2h	○
TRF × χ 二乗 × GA	0.842	25h	○
TRF × ANOVA × GA	0.83	25h	○

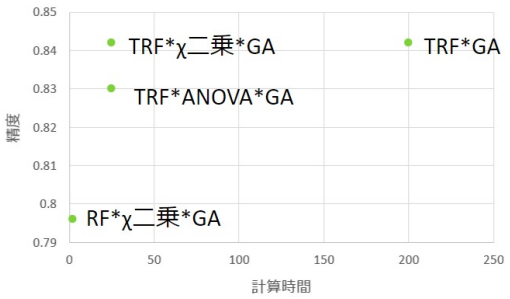

図 2 交互作用を考慮できるモデルににおける, 精度と実行時間に関するグラフ

表 1 より, ①精度の改善には GA が, ②計算時間の短縮にはχ二乗フィルタが, ③交互作用の考慮には GA が有効であり, 3 つの観点から提案手法であるハイブリッド特徴選択を用いた分類の有効性が示された．また, 本研究における, χ二乗検定と GA を用いたハイブリッド特徴選択と TRF の組み合わせの有効性を, 他のハイブリッド特徴選択あるいは分類モデルと比較することで示す．比較結果を表 2 に示す.

さらに, 交互作用を考慮できるモデルに対し, 縦軸に精度を, 横軸に計算時間を示したものを図 2 に示す．交互作用が考慮できるもののうち, 精度が高くかつ計算時間が短いモデル, すなわち図 2 における左上に位置するモデルにあたる提案手法が, 3 つの観点から総合的に評価して課題に適しているということが示された.

5 まとめ

本研究では, χ二乗と GA を利用したハイブリッド特徴選択を用いた, TRF による大豆の収量分類モデルを構築した．提案モデルでは, χ二乗フィルターの利点である次元削減による計算時間の短縮により GA の欠点である計算時間に対処した．さらに, GA の利点である交互作用の考慮によりχ二乗フィルターの欠点である交互作用に関する情報欠損に対処した．これらのハイブリッド特徴選択により互いの欠点を補いあい, 分類精度の改善を達成につなげることができた．しかし, 機械学習モデルはその解釈の難しさから, 農家の方にはハードルが高く受け入れがたいことが多い．今後はこれらの結果と実際の知見をもとに, 解釈付けしていくことが重要である.

また, 今回は全国の圃場に対し一元的にモデル構築を行ったが, 今後は地域特性や圃場特性も鑑み, クラスタごとの特徴選択, さらにはクラスタ分類と特徴選択の結果から収量に関する知識抽出を目指す.

謝辞

農林水産省「収益力向上のための研究開発 (中課題番号：15653568，中課題名：多収阻害要因の診断法及び対策技術の開発)」の支援により実施した.

参考文献

[1] Guido Sciavicco,Ionel Eduard Stan,”Knowledge Extraction with Interval Temporal Logic Decision Trees”,27th International Symposium on Temporal Representation and Reasoning (TIME2020),Vol178,pp.9:1-9:16(2020)

[2] Fateme Moslehi1 · Abdorrahman Haeri1, ”A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection”, Journal of Ambient Intelligence and Humanized Computing,Vol11,pp1105-1127(2019)

[3] 国立研究開発法人農業・食品産業技術総合研究機構，メッシュ農業気象データシステム， https://amu.rd.naro.go.jp/（最終閲覧日: ２０２３年６月）

[4] Midori Namba, Kouhei Umejima, Ryo Nishide, Takenao Ohkawa, Seiichi Ozawa,Noriyuki Murakami and Hiroyuki Tsuji, ”Optimal Pattern Discovery based onCultivation Data for Elucidation of High Yield Inhibition Factor of Soybean”,Proceedings of the 5th IIAE International Conference on Intelligent Systemsand Image Processing