

大豆の遺伝特性と環境特性に基づく Temporal Random Forest を利用した
混合モデルによる収量予測Yield Prediction by Mixture Model Using Temporal Random Forest
Based on Genetic and Environmental Characteristics of Soybeans大橋真琴¹⁾ 大川剛直¹⁾
Makoto Ohashi Takenao Ohkawa

1 はじめに

近年、農業では人口減少下においても生産力を維持できる生産性の高い農業を実現するため、スマート農業の実用化が進められている。その1つとして、機械学習による作物の収量予測が挙げられる。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。また我々の研究グループでは、日本の食文化を支える大豆という作物に注目し収量予測を行っている。本研究では、大豆の収量予測モデルの精度改善を目指し、収量に影響を及ぼす特徴を、品種などの遺伝特性と気象条件などの環境特性の2つに分けて考え、両方を考慮したモデルを提案する。

2 提案手法

2.1 遺伝特性と環境特性

作物の生育には、品種による遺伝特性と生育環境の気象条件などの環境特性が複雑に絡み合って作用している。同一の環境で育ったとしても、その影響は品種によって大きく異なる。また、品種によって注目すべきポイントや収量のポテンシャルが異なる。よって次項に示す混合モデルを提案する。

2.2 混合モデル

図1に提案モデルの概要を示す。提案する混合モデルでは、遺伝特性によりデータをグループ化し、グループ毎に環境要因を利用した予測モデルを構築し、混合させる。

はじめに、遺伝特性に基づきデータを条件ごとに2つのクラスタに分類する。条件がD項目の場合、各データはそれぞれの項目でどちらかのクラスタに分類され、D個のクラスタに所属することになる。次に、クラスタ毎に収量予測モデルを構築する。ここでは次項で説明するTemporalRandomForestにより回帰を行う。これにより1つのデータに対しD個の予測値が生成される。最後に、D個の予測値から加重平均を求め、これをデータに対する混合モデルの予測値とする。

これにより、遺伝特性を考慮した上で環境特性に基づく収量予測が可能になる。また、複数のクラスタに分けるのではなく項目ごとに2分しデータを複数のクラスタに所属させ予測を行うことで、クラスタ毎のデータ数が一定となる。そのため教師あり学習に必要なデータ数が担保できる。さらに、マジョリティなデータに対しては遺伝特性が異なるマイノリティなデータによる影響を抑えることができる。対してマイノリティなデータには、遺伝特性から類似の傾向を示すマジョリティなデータの学習を利用し予測することが可能となる。

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

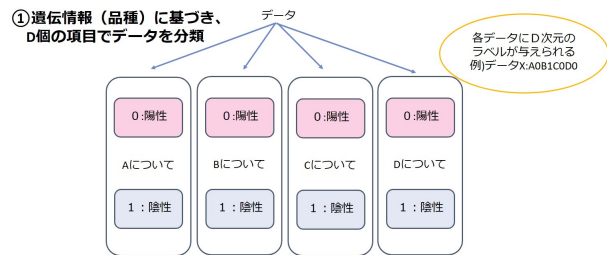


図1 提案モデルの概要

2.3 Temporal Random Forest

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予測が可能であるが、モデルが複雑であるため、学習に時間がかかってしまう。

3 実験

実大豆圃場における大豆の品種データ、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、栽培品種、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。

3.1 データの前処理

環境特性を利用した予測モデル構築では、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。

3.1.1 大豆の栽培特性に基づくデータ加工

はじめに、大豆の栽培特性に基づき2つのデータ加工を行う。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]。

3.1.2 GAを用いたハイブリッド特徴選択

特徴選択手法は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^2 乗検定と遺伝的アルゴリズム (GA) を使い、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。

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。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。

GAによる特徴量選択では、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]。

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②計算時間の短縮③交互作用を考慮したモデルが実現する。

3.2 実験結果

実験では、表1に示すよう、4つの分類条件からクラスタを作成した。

表1 遺伝特性によるクラスタ分類条件

	分類条件
A	倒れやすい品種か
B	連作障害のある品種か
C	多収品種か
D	大粒品種か

提案手法と他の予測モデルとの比較結果を表2に示す。各実験において、10分割交叉検証を行いその平均平方二乗誤差を求めた。

表2 提案モデルと比較手法による大豆の収量予測結果

	予測誤差 (平均平方二乗誤差)
RF	84
TRF	77
提案手法	70

また、遺伝特性に基づくクラスタごとの予測と混合モデルによる予測の比較を表3に示す。ここで、表1の条件に対し、該当するものを0・該当しないものを1とした。すなわち、A0は倒れやすいクラスタに対するTRFによる収量予測の結果である。また考察のため、各モデルにおいて、多収データのみ・低収データのみに対して実験を行った。

表3 遺伝特性に基づくクラスタごとの予測と混合モデルによる予測

	予測誤差	多収データの誤差	低収データの誤差
A0	75	76	75
A1	77	77	77
B0	75	74	77
B1	77	76	77
C0	71	70	75
C1	74	75	74
D0	74	73	75
D1	78	79	77
混合モデル	70	69	70

3.3 考察

表2,表3の結果から、今回の提案手法では他の手法と比べ、大豆の収量予測に対し誤差を抑えることができた。これは、遺伝特性により環境特性の影響度が異なるため、クラスタ毎に予測モデルを構築した結果改善されたと考えられる。また、表3より、クラスタごとの分類にとどめず混合モデルとしたことで、分割によるデータ数減少による精度の低下を防ぐことができた。さらに、多収データ・低収データに偏らない予測を実現した。

4 まとめ

本研究では、収量に影響を及ぼす特徴を、品種などの遺伝特性と気象条件などの環境特性の2つに分けて考え、遺伝特性によりデータをグループ化した混合モデルにより大豆の収量予測を行った。予測には環境特性の時系列性からTemporal Random Forestを利用し、その計算量を抑えるためにGAを用いたハイブリッド特徴選択を利用した。実験の結果から、この混合モデルにより、収量予測精度の改善が見られた。今後は実地利用に向け、さらなる精度の向上と、予測モデル構築の自動化を目指す。

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:多収障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 Sciacicco, 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 on Cultivation Data for Elucidation of High Yield Inhibition Factor of Soybean", Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing