

機械学習を用いた大豆の収量分類および収量に寄与する知識抽出

大橋真琴[†] 大川剛直[†]神戸大学大学院システム情報学研究科[†]

1. はじめに

大豆の栽培管理手法確立のために、収量に寄与する知識の抽出が望まれている。本研究では利用するデータの時系列性から特徴量間の交互作用を考慮し、大豆の収量分類および特徴評価により知識の抽出を行うモデルを提案する。図1に示すように、提案モデルでは①大豆の栽培特性に基づくデータ加工②Genetic Algorithm(以下、GA)による特徴選択③Temporal Random Forest(以下、TRF)による収量分類④バスケット分析による特徴評価から構成される。

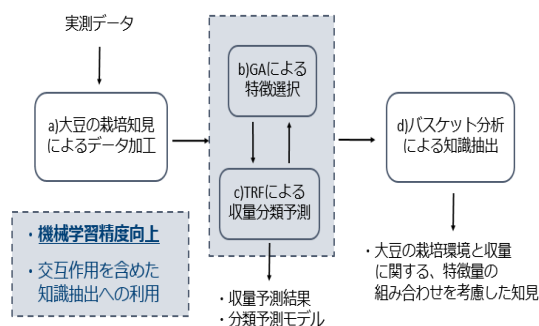


図1: 提案モデルの概要

2. 大豆の栽培特性に基づくデータ加工

大豆は屋外で栽培される作物であり、その生育は気象に大きく影響を受ける。そこで本研究では大豆の栽培環境データとして気象条件を利用する。このとき、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栽培環境データに対し2種類の加工を行う[1]。1つめはデータの取得基準を大豆の生育ステージ毎とするステージ分割である。もう1つは平年値と比較し表現することで多変量データの粒度をそろえる定性値化である。

3. GAによる特徴選択

特徴選択は機械学習の精度を向上させる手法の1つである。また、本研究で知識抽出に利用するバスケット分析では特徴量のサブセットの比較を行う。特徴量が n 次元の場合、サブセットの総数は 2^n となる。そこで、GAによる探索の過程を利用することで効率的なサブセットの生成を行う。GAによる特徴選択の流れを以下に示す。ここで、遺伝子は各特徴量に対しサブセットに含む場合は1を、含まない場合は0を与えたものとする。

STEP 1: 第1世代の遺伝子をランダムに生成する。

STEP 2: 遺伝子が示すサブセットに対しTRFを用いて分類予測精度を求める。

STEP 3: 交叉・突然変異により次の世代を作成する。

STEP 4: STEP2, 3を繰り返す。

本研究ではTRFを用いた分類予測精度をGAの適応度とし、GAによる探索は予測精度が向上する方向に行われる。よって最終世代の遺伝子が特徴選択結果である。また、GAによる探索の過程で得られるサブセットと分類予測精度を知識抽出に用いる。

4. TRFによる収量分類

本研究で扱うデータが多変量時系列データであるという特性を活かし、Random Forestを拡張したTRF[2]を用いて収量が低収であるか、多収であるかの分類予測を行う。先ほどGAによる特徴選択の流れで述べた通り、GA遺伝子が示す特徴サブセットを用いて分類予測を行い、その精度を適応度として返す。

「Soybean yield classification and knowledge extraction contributing to yield using machine learning」

[†]「Ohashi Makoto, Ohkawa Takenao, Kobe University」

5. バスケット分析による特徴評価

バスケット分析はマーケティングに利用される分析手法である。本研究ではバスケット分析を用いて収量分類予測精度の高い特徴サブセットの傾向を分析することにより、収量に寄与する気象条件に関する知識抽出を行う。ここでは支持度とリフト値という2つの指標を用いて特徴評価を行う。支持度は特徴量あるいは特徴量組み合わせの出現率を示し、支持度が高いものは収量への寄与度が高いといえることができる。リフト値はある特徴量が他の特徴量へ与える影響を示す指標であり、特徴量どうしが組み合わせることによって生じる交互作用の存在を確認することができる。リフト値が高い組み合わせに対してはまとまりとしてみるのが重要である。

6. 実験

全国の大豆圃場から収集した栽培データと農研機構メッシュ農業気象データシステム[3]が提供する気象データを利用し実験を行う。2015年から2018年に収集された160圃場455データを利用する。また、栽培環境を表す気象として平均気温・降水量・風量・気温差の4条件を利用する。データ加工では大豆の生育段階に則り13のステージ分割を行い、気象条件は平年値と比較し3段階で、収量は250kg/10aを閾値とし低収・多収の2値で表現する。提案手法で利用したTRFとGAの組み合わせと、他の分類器との収量分類予測の精度を表1に示す。本実験では予備実験に基づきGAの1世代の遺伝子数を70、世代数を100とした。

表1：予測器の比較結果

	SVM	SVM*GA	RF	RF*GA	TRF	TRF*GA
accuracy	0.733	0.755	0.788	0.800	0.837	0.841
F値	0.698	0.724	0.746	0.776	0.810	0.822

ここから、TRFとGAの組み合わせが最も利用するデータの特徴を捉えた分類ができていることがわかる。また、様々な特徴選択アルゴリズムを用いた場合のTRFによる分類予測の結果を表2に示す。

表2：特徴選択アルゴリズムの比較結果

TRF*	ランダム サンプリング	X二乗 フィルター	PSO	GA	ブルート フォース (wrapper)
accuracy	0.837	0.803	0.835	0.838	0.841
F値	0.810	0.781	0.805	0.819	0.822

(ブルートフォースは計算時間が現実的ではなく結果が得られず)

7. まとめ

本稿では、①大豆の栽培特性に基づくデータ加工②GAによる特徴選択③Temporal Random Forestによる収量分類④バスケット分析による特徴評価の4段階からなる大豆の収量分類および収量に寄与する知識抽出モデルを提案した。実験結果からGAとTRFの組み合わせによる収量分類予測は他の機械学習手法と比較し精度が向上する結果となった。また、特徴選択手法としてGAを用いることは、精度の向上に加え、計算時間と交互作用の考慮という両面で効果的であると言え、本研究の主目的である知識抽出の準備として適切であると考えられる。今後はこれらの結果をもとにバスケット分析による知識抽出結果と大豆栽培あるいは気象の知識を照らし合わせることで知識抽出を目指す。

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] Midori Namba, Kohei Umejima, Ryo Nishide, Takenao Ohkawa, Seiichi Ozawa, Noriyuki Murakami, Hiroyuki Tsuji, "Optimal pattern Discovery to Reveal the High Yield Inhibition Factor of Soybeans", Journal of the Institute of Industrial Applications Engineers Vol.6, No.2, pp.66-72, (2018)
- [2] 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).
- [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/>