Vol.2018-MPS-119 No.15
2018/7/30

機械学習を用いた水稲の収量予測について

前田 佑一郎[1]　五葉谷 太一[2]　西内 俊策[3]　北 栄輔[1]

概要：本稿では，水稲の収量予測について述べる．各圃場の気象情報，栽培情報，位置情報を説明変数として，XGBoost により収量予測モデルを作成する．最も予測精度が高かったのは，気象情報から説明変数を決定するときに積算区間を 2 つ (田植え日から出穂日，出穂日から登熟日) に分ける手法であり，その予測精度は 74.4%であった．また収量予測における変数重要度についても評価した結果，気象情報は収量予測において影響が大きいことがわかった．

キーワード：機械学習，XGBoost，水稲．

Yield prediction of paddy rice with Machine Learning

Yuichiro Maeda[1]　Taichi Goyotani[2]　Shunsaki Nishiuchi[3]　Eisuke Kita[1]

Abstract: In this paper, the yield prediction of paddy rice is defined by use of XGBoost from the weather information, the cultivation data and the location information of the paddy rice field. The best accuracy is estimated as 74.4% when the weather information is integrated at two intervals such as planting date to heading date and heading date to ripening date. The discussion on the variable importance of explanatory variables for the prediction accuracy revealed that the weather information was very effective in yield prediction.

Keywords: Machine learning, XGBoost, Rice cultivation.

1. 緒論

日本の農業は，総産出額が 1984 年に 11 兆 7 千億円に達して以降，コメを中心に減少傾向にある [1], [2]．その原因に農業就業人口の減少や篤農家の高齢化などがある．これらに加えて，近年の地球温暖化による異常高温化に伴い，収量が低下しているという報告がある [3], [4], [5]．これらの問題を解決するために，データサイエンスや情報科学の農業分野への導入が広く行われている [6], [7].

本研究の目的は，水稲の収量予測モデルの設計である．目的変数として水稲の収量を，説明変数として気象情報，栽培情報，位置情報を用いる．気象情報には日最高気温，日最低気温，日照時間があり，田植え日から登熟日までの区間をいくつかの部分積算区間に分けて求めた積算値を用いる．栽培情報には播種様式や栽植密度などを用いる．アルゴリズムとして XGBoost を用いて節三井変数から目的変数を予測するモデルを作成する [8], [9], [10].

本論文の構成は以下のようになっている．第 2 では提案アルゴリズムについて述べる．第 3 では実験結果を示す．第 4 はまとめである．

2. 予測アルゴリズム

2.1 水稲の栽培ステージ

水稲発芽から収穫までのステージは育苗期，分げつ期，幼穂発育期，登熟期に分かれる．育苗期とは，たねもみの発芽から田植えに至るまでの段階をさす．通常，日本では水田ではなく育苗器やビニルハウスを用いて，20 から 30 日間で育苗を行う．分げつ期では，茎がいくつにも分かれる．分げつは収量に影響を与える．幼穂発育期とは，イネが最後の葉であを作り，穂の元である幼穂を作る．幼穂の

[1] 名古屋大学大学院情報学研究科, Nagoya 464-8601, Japan
[2] 名古屋大学大学院情報科学研究科, Nagoya 464-8601, Japan
[3] 名古屋大学大学院生命農学研究科, Nagoya 464-8601, Japan

形成は出穂期の約 1 ヶ月前に始まる．この時期を幼穂発育期という．出穂後，稲は登熟期を迎える．

2.2 勾配ブースティング

XGBoost は勾配ブースティングの一種である．勾配ツリーブースティングでは，ツリーアンサンブルモデルのリーフの重みを更新して，評価式を最小化できる最適なモデルを導出する．その概要として，まずツリーアンサンブルモデルの予測値は，以下のように算出される．

$$\hat{y_i} = \phi(x_i) = \sum_{k=1}^{n} K = f_k(x_i), f_k \in \mathcal{F} \qquad (1)$$

ただし，

$$\mathcal{F} = f(x) = w_{q(\mathbf{x})}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T) \qquad (2)$$

は回帰木空間である．x_i は入力，$\hat{y_i}$ は出力，q はツリーの構造，T はツリー内のリーフの数である．各 f_k は独立したツリー構造 q とリーフの重み w に一致する．w_i は i 番目のリーフのスコアを表す．この予測値は，

$$L(\phi) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k) \qquad (3)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \qquad (4)$$

で評価できる．l は予測値 $\hat{y_i}$ と目標値 y_i との差を求める損失関数である．モデルの複雑さを表す Ω は正則化項であり，過学習を避けるために重みを平滑化する働きがある．

2.3 目的変数と説明変数

研究の目的は水稲の収量予測である．稲作において，収量に影響を与える要因の一つは気象情報である．また，栽植密度などの栽培情報も収量に関与していると考えられる．そこで，目的変数として水稲の収量を採用し，説明変数には気象情報と栽培情報，これらに加えて，試験場ごとの位置情報を説明変数に加える．

気象情報については，最高気温，最低気温，日照時間があり，これらを田植え日から登熟日までの区間を 1 つまたは複数の区間に分けて積算した値を用いる．積算区間の影響を評価するために，3 種類の区間を比較する．第 1 は，田植え日から登熟日までを 1 区間とする場合である．第 2 は，田植え日から出穂日と出穂日から登熟日の 2 区間とする場合である．それらを”max_tmp_1”と”max_tmp_2”とする．第 3 は，4 区間に分けた場合である．それらを”max_tmp_1”，”max_tmp_2”，”max_tmp_3”，”max_tmp_4”とする．それぞれを Test 1，Test 2，Test 3 とする．

栽培情報については，年度，播種様式，移植方法，施肥水準，基肥量，追肥の回数，追肥の量，穂長，稈長，栽植密度，育苗日数，出穂日数，登熟日数の 13 種類の栽培データがある．

位置情報については，試験地の緯度と経度がある．

表 1　収量予測結果 (2011-2014)
Table 1　Yield prediction results (2011-2014)

category	Test 1	Test 2	Test 3
Average	7.35%	7.25%	7.01%
Median	6.16%	6.41%	5.94%
Max	26.53%	29.83%	27.13%
N	614	614	614
N_over 10%	150	135	144
N_over 20%	17	22	14
Acc	72.8%	74.4%	74.3%

3. 解析結果

3.1 問題設定

実際の農業現場では，収量の予測は誤差 10%以内であることが求められている．したがって，本実験では，予測誤差 10%以内のデータを正解データとして定義する．また予測精度は全データのうち 10%以内で予測できたデータの数を算出して求めることとする．予測精度は次式で定義される．

$$Accuracy = \frac{T}{T+F}\ (\%) \qquad (5)$$

ここで，T は誤差 10%以内で予測できたデータの数，F は誤差 10%以上で予測したデータの数を表す．

3.2 実験結果

Table 1 に各 Test の予測結果とその内訳を示す．ここで，”Average”は収量予測における平均誤差，”Median” はその中央値を, ”Max”はその最大誤差を示す．”N”は全データ数を，”N_over 10%”は予測誤差 10%以上で予測を行なったデータ数を，”N_over 20%”は予測誤差 20%以上で予測を行なったデータ数示す．”Acc”は全データのうち正しく予測できたデータ数から予測精度を算出した値である．

収量の予測精度は 10%以内が求められている．予測精度が最も高かったのは Test 2 である．Test 1 と Test 3 は average，median，max の精度が Test 2 よりも優れていることがわかる．Test 2 は最も良い予測精度であるが，予測誤差 20%以上で予測してしまったデータ数が最も多いのも Test 2 のモデルである．

各テストの詳細な予測結果を，それぞれ Table 2，3，4 に示す．最も優れた予測モデルは Test 2 である．しかし，median を見てみると，2011-2014 年の全てを通して，Test 3 のモデルが優れておることがわかる．また，max についても，Test2 のモデルは Test 1,Test 3 比較して全体的に精度は良くない．全体の精度は Test 2 が一番優れているが，予測誤差の割合や最大誤差などを精査すると，Test 1，Test 2 の予測モデルの方が優れていると言う興味深い結果となった．また，全体を通して，2011,2012 年と比較して，2013，2014 年の予測精度が低いことがわかる．これについ

表 2　収量予測結果 (2011-2014, Test 1)
Table 2　Yield prediction results (2011-2014, Test 1)

year	2011	2012	2013	2014
Average	7.24%	6.73%	7.42%	7.34%
Median	6.12%	5.97%	6.19%	6.16%
Max	21.67%	17.75%	26.54%	26.51%
N	164	146	135	169
N_over 10%	32	41	31	46
N_over 20%	2	0	7	8
Acc	79.27%	71.92%	71.86%	68.05 %

表 3　収量予測結果 (2011-2014, Test 2)
Table 3　Yield prediction results (2011-2014, Test 2)

year	2011	2012	2013	2014
Average	6.93%	6.86%	7.25%	7.29%
Median	6.25%	6.25%	6.41%	6.46%
Max	24.41%	19.29%	29.83%	23.801%
N	164	146	135	169
N_over 10%	35	31	26	43
N_over 20%	6	0	10	6
Acc	75.0%	78.77%	73.33%	71.01 %

て，参考文献 [11] によると，2013 年度は 4 月中旬〜5 月上旬で全国的に低温傾向となったものの，その他の時期で暖気に覆われ，気温変動が大きかった．また，2014 年は 3〜5 月にかけての日照時間が統計を開始した 1946 年以降最も多かったことに加え，3 月下旬と 5 月下旬に南から暖気が流れ込み，気温が平年を大幅に上回ったと報告されている．例年と異なる気象情報の変化と本実験での気象情報の分割方法ではその変化を適切にとらえきれなかった可能性が考えられる．

本実験で Test 2 の予測モデルの予測精度が最も優れていた要因は，予測精度が低い 2013,2014 年にあると考えられる．N_over10%が Test 1 と Test 3 はそれぞれ 77,87 であるのに対して，Test 2 は 69 である．結果として全体の予測精度を高めたと考えられる．しかし，実際の農業現場においては，全体の精度の良さを重視するのか予測誤差が大きいデータを避けることを重視するのかで，モデルの評価の仕方は変わってくる．以上の分析より，温度と日照時間の積算期間を分割して説明変数に加えたほうが，予測の精度が向上した．つまり，積算期間を分割することにより，各データの気象情報の細かい変位まで考慮して予測が可能になったと考えられる．本実験では積算区間の分け方を 3 パターンに分けて評価した．その結果 Test 2 の積算期間の分割方法が最も良い精度を残したが，この他にも様々な積算区間の分割方法を試し，収量予測に最も効果的な積算方法を調査する必要がある．

表 4　収量予測結果 (2011-2014, Test 3)
Table 4　Yield prediction results (2011-2014, Test 3)

year	2011	2012	2013	2014
Average	7.01%	7.03%	6.72%	6.96%
Median	5.92%	5.94%	5.54%	5.92%
Max	27.13%	20.26%	27.13%	23.44%
N	164	146	135	169
N_over 10%	30	27	37	50
N_over 20%	2	1	7	4
Acc	80.49%	80.82%	67.4%	68.05 %

表 5　変数重要度ランキング (Test 1)
Table 5　Variable importance ranking (Test 1)

Rank	Variable	Importance
1	max_tmp	0.12
2	sunshine_time	0.10
3	min_tmp	0.097
4	Rice plant height	0.088
5	Year	0.087
6	Latitude	0.073
7	Days from planting to heading	0.068
8	Rice ear length	0.067
9	Longitude	0.051
10	Days from heading to ripening	0.043

表 6　変数重要度ランキング (Test 2)
Table 6　Variable importance ranking (Test 2)

Rank	Variable	Importance
1	sunshine_time_1	0.093
2	min_tmp_1	0.088
3	min_tmp_2	0.084
4	sunshine_time_2	0.082
5	max_tmp_1	0.077
6	max_tmp_2	0.075
7	Rice plant height	0.073
8	Year	0.061
9	Rice ear length	0.054
10	Latitude	0.053

表 7　変数重要度ランキング (Test 3)
Table 7　Variable importance ranking (Test 3)

Rank	Variable	Importance
1	Rice plant height	0.086
2	sunshine_time_1	0.063
3	sunshine_time_2	0.060
4	sunshine_time_4	0.058
5	sunshine_time_3	0.057
6	min_tmp_1	0.050
7	max_tmp_4	0.050
8	min_tmp_4	0.049
9	min_tmp_2	0.048
10	min_tmp_3	0.048

3.3 説明変数の重要度

各予測モデルにおける変数重要度のランキングの上位 10 変数を Table 5，6，7 に示す．

Test 1，Test 2，Test 3 の全予測モデルにおいて，気象情報の重要度が大きいことがわかる．その中でも，気象情報の積算区間を分割した Test 2,Test 3 を見ると，Test 2 では田植え日から出穂日までの日照時間を示す sunshine_time_1 が 1 位に，また Test 3 では，田植え日から出穂日の期間の中でも前半の日照時間が収量に強く影響を及ぼしていることがわかる．このことから，日照時間の中でも田植え日から出穂日までの日照時間の積算が収量予測において重要であることがわかる．気温情報においては，Test 1 では最高気温の積算の説明変数の重要度が高い一方で，Test 2, Test 3 では最低気温の積算の説明変数の重要度が高いという結果になった．栽培データの中では，year や the height, the length などの重要度が高い．また，位置情報の中では，latitude が変数重要度の上位としてランクインしている．longitude より latitude の重要度が高い要因として，緯度の違いにより気候や日射量は大きく変わるためであると考えられる．

4. 結論

本実験では，水稲の収量予測モデルの作成を行なった．予測モデルの作成には，XGBoost を用いて，説明変数には気象情報，栽培情報，位置情報の 3 種類の情報を用いた．気象情報を用いる際，田植え日から登熟日までの積算情報を用いたが，その積算する区間を変更し，収量予測への影響を評価した．最も予測精度が高かったのが積算区間を 2 つ (田植え日から出穂日，出穂日から登熟日) に分ける手法であり，その予測精度は 74.4%であった．また収量予測における変数重要度についても評価した結果，気象情報は収量予測において影響が大きいという結果になった．

謝辞 本研究は生研支援センター「「知」の集積と活用の場による研究開発モデル事業」の支援を受けて行った．また，研究に用いた水稲の栽培情報は愛知県農業総合試験場より提供を受けた．ここに記して感謝の意を表する．

参考文献

[1] Ministry of Agriculture. “Annual agricultural production output and production agricultural income.” http://www.e-stat.go.jp/SG1/stat/List.do. (2008) (in Japanese) 2017/12/24 accesced.
[2] National Agricultural Cooperative Association. “Think about Japanese ingredients 2013. Current situation of agriculture in Japan.” https://www.zennoh.or.jp/japan_food/02.html. (in Japanese) 2017/12/24 accesed.
[3] Wakamatsu, Ken-ichi. “Effects of high air temperature during the ripening period on the grain quality of rice in warm regions of Japan.” Bulletin of the Kagoshima Prefectural Institute for Agricultural Development. Agricultural Research (in Japanese). (2010).
[4] Hasegawa, Toshihiro, et al. “Recent warming trends and rice growth and yield in Japan.” MARCO Symposium on Crop Production under Heat Stress: Monitoring, Impact Assessment and Adaptation. National Institute for Agro-Environmental Studies, Tsukuba, Japan. 2009.
[5] Okada, Masashi, et al. “A climatological analysis on the recent declining trend of rice quality in Japan.” J. Agric. Meteorol 65.4 (2009): 327-337.
[6] Horie, T., et al. “The rice crop simulation model SIMRIW and its testing.” Modeling the impact of climate change on rice production in Asia (1995): 51-66.
[7] Wakiyama, Yasuyuki, Kimio Inoue, Kou Nakazono. “A simple model for yield prediciton of rice based on vegetation index derived from satellite and AMeDAS data during ripening period.” Journal of Agricultural Meteorology (in Japanese) (2003).
[8] Scalable and Flexible Gradient Boosting. https://xgboost.readthedocs.io/en/latest//. 2017/12/24 accessed.
[9] Jain, Aarshay. “Complete guide to parameter tuning in Xgboost.” (2016).
[10] Chen, Tianqi, Carlos Guestrin. “Xgboost: A scalable tree boosting system.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, (2016). 785-794.
[11] JMA/Climate system monitoring annual report. https://www.data.jma.go.jp/gmd/cpd/diag/nenpo/. (in Japanese) 2017/12/24 accesed.