「マルチメディア，分散，協調とモバイル
(DICOMO2021)シンポジウム」 令和3年6月-7月

栽培データの分布不均衡性を考慮した植物生理状態推定の検討

藤浪一輝[1] 大石直記[2] 二俣翔[2] 峰野博史[3]

概要： 農業従事者の不足や技術の消失といった課題に対して，スマート農業の実現に向けた取り組みが行なわれている．スマート農業の一環として栽培管理支援システムの開発が進められており，光合成速度や蒸発散速度といった植物の生理状態を表す指標を推定する技術が期待されている．植物生理状態の既存の推定手法は設置コストの高さや栽培管理の妨げなど実用面での弊害が存在するため，本研究では低コストかつ非接触なセンサデータから機械学習で推定する手法を検討する．機械学習を用いる場合の問題点として栽培データの不均衡性問題があり，推定精度の低下が危惧されるため，リサンプリング処理を適用しデータ分布を変化させることで推定精度の向上を図る．ただし，回帰問題に対する不均衡性は議論されることが少なく，適用可能な既存のリサンプリング手法は多くない．また，既存手法では目的変数の不均衡性のみに着目しており説明変数の分布を考慮していないため，一部の環境条件の情報を失う可能性がある．そこで，各環境条件の情報を保持したまま不均衡性を解消するリサンプリング手法としてCREAMER(**C**lustering-based **RE**s**A**mpling **ME**thod for **R**egression)を提案する．CREAMER は目的変数と説明変数の両方をクラスタリングし，各クラスタにリサンプリングを適用することでクラスタ間のデータ数が均等になるように変換される．イチゴの光合成速度と蒸発散速度について既存手法と CREAMER を適用した際の推定精度の比較検証を行ったところ，既存手法よりも高精度に推定できることを確認でき，リサンプリングを行わない場合の結果と比較して光合成速度では MAE を 6.25%，RMSE を 9.38%，蒸発散速度では MAE を 53.11%，RMSE を 40.74%削減できた．

Estimation of plant physiological state considering imbalanced cultivation data distribution

KAZUKI FUJINAMI[1] NAOKI OISHI[2] NATSURU FUTAMATA[2]
HIROSHI MINENO[3]

1. はじめに

近年，国内の農業分野では新規農業者の高い離農率が課題とされており，それに伴う農業従事者の不足と高齢化，熟練農家が持つ専門技術・技能の継承困難性が問題視されている．多くの農業従事者は長期にわたる経験の中で試行錯誤を繰り返し，暗黙知として技術を獲得し安定した栽培を可能としているため，十分な栽培技術を持ち合わせていない新規農業者が，持続的に安定した栽培をすることは困難であり,収支が成り立たず農業の継続困難に及んでいる．こういった問題を解決するため，スマート農業[1]の一環として農家の栽培支援を行う栽培管理支援システムの開発が行なわれている．そこで，植物の周辺情報から植物の生理状態を推定する手法の確立が期待されている．

本研究では，植物生理状態を示す指標として光合成速度と蒸発散速度を機械学習で推定する手法を検討する．光合成速度を選択した理由は，一般的に植物が光合成活動により有機物を合成し内部構造を発達していくため，植物の成長速度を把握する指標として有効だからである．蒸発散速度を選択した理由は，植物は過剰灌水を行うと病害発生リスクが高まり，特に果菜類においては果実の裂果や低糖度化といった作物品質の低下を招くため，植物体内水分量に応じて変化する蒸発散速度は適切な灌水の制御を行う指標として有効とされるからである．これらの指標の計測方法は既に存在するものの，実用化に対する課題があり一般農家の普及には至っていない．また，植物の栽培データは不均衡な分布といった機械学習に対して推定値や予測値の精度低下を招く課題を持つ．そこで，本研究ではデータの不均衡性を解消する前処理手法として，CREAMER (**C**lustering-based **RE**s**A**mpling **ME**thod for **R**egression)というデータのリサンプリング手法を提案する．

以降，第 2 章では光合成速度及び蒸発散速度の従来の計測手法と課題，第 3 章では関連研究について述べる．第 4 章では提案手法のアルゴリズムについて説明し，第 5 章では栽培データでの検証実験結果，第 6 章で本稿のまとめを述べる．

1 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
2 静岡県農林技術研究所
Shizuoka Prefectural Research Institute of Agriculture and Forestry
3 静岡大学学術院情報学領域
College of Informatics, Academic Institute, Shizuoka University

2. 従来の計測手法と課題

2.1 光合成速度

光合成速度の代表的な計測手法[2][3]として，チャンバー

© 2021 Information Processing Society of Japan

（同化箱）法が用いられる．チャンバー法は，チャンバーと呼ばれる箱型の透明な容器で植物体を覆うように設置し，内部に取り付けたセンサで収集したCO_2濃度と気温から光合成速度を求める手法である．チャンバー下部の吸気部と上部の排気部に CO2 センサを取り付け，光合成前後の CO2 濃度差とチャンバー内気温から光合成速度を算出する．チャンバー法は外部と空間を区切るため，人の出入りや窓の開閉による気温や風の変化を受けづらく高精度な計測が可能である．しかし，他の植物体との成長環境に差が生じることから，個別に管理する必要があり栽培が難化する恐れがある．また，チャンバーは一般的に大掛かりかつ高価な機材であり，農場の規模に比例して設置コストと経済的コストが増加するため，実用的な利活用には課題が残る．

2.2 蒸発散速度

蒸発散速度の代表的な計測手法[4][5]として，土壌水分減少法，重量法および光合成速度と同様のチャンバー法が用いられる．土壌水分減少法，重量法はテンシオメーターやロードセルなどの計測機器で灌水前後における土壌と植物体の重量，水分量の変化量を計測する手法である．蒸発散により失った水分を直接計測することで蒸発散速度を算出できるが，土壌水分量は空間的に不均一なため正確な計測が困難である．不均一な水分分布の土壌に対するセンサ設置の代表地点を選出する手法は提案されている[6]が，選出する過程で複数箇所の土壌水分量を計測する必要があり，現状は培地の複数地点にセンサを設置して平均水分量を計測するしかない．従って，チャンバー法と同様に設置コストと経済的コストの増加が考えられる．さらに，重量法は蒸発散の微小な重量変化を捉えるため，風や農作業による揺れが大きなノイズとなり計測に支障を生じやすい．

3. 関連研究

第 2 章で述べた課題から，設置コストが小さく，且つ農作業によるノイズが発生しにくいセンサデータを用いて，機械学習により推定する手法を検討する．しかし，植物の栽培データは不均衡な分布を持っており機械学習に適していないことが明らかとなってきた．そこで，分布不均衡性を解消するリサンプリング手法に着目し，推定精度の向上を図る．

3.1 分布不均衡性

機械学習は与えられたデータから網羅的にパラメータ探索を行うため，学習データの分布が不均衡だと偏った学習モデルが構築される．また，不均衡データでは希少な値への推定精度が重要な場合が多い．植物生理状態推定においても，生産的活動が最大となる昼間の推定が重要だが，半日以上を占める日没以降の推定はあまり重要でない．

不均衡データの解決法には大別して，アルゴリズムレベルの手法とデータレベルの手法がある．アルゴリズムレベルの手法は，モデルの学習段階で不均衡性を考慮する方法で，データごとにコストの与え方を変動させるコスト考慮型学習や，複数の弱学習器を用いたアンサンブル学習によりバイアスの低減を図る．しかし，データや学習器に対する高度な理解を必要とし，新しいデータに対しても逐次修正が必要になるという課題がある．

一方，データレベルの手法は，不均衡データの少数グループに対して新しいデータを生成(オーバーサンプリング)，または多数グループに対してデータの除去（アンダーサンプリング）を行い，均衡な分布に変換する処理を行う方法である．課題としては，リサンプリングするデータの対象を誤ると，重要なデータの除去や多量の同一データが生成され精度が低下する恐れがある． ただし，モデルの学習と独立しており，データのみに配慮すればよいため，精度低下時の問題の切り分けを実施しやすいという利点がある．

分布不均衡性の問題はクラス分類問題で提起されることが多いが，本研究が対象とする回帰問題でも同様な課題に直面しているといえる．既存のリサンプリング手法はそれほど多くなく[7]，またデータセットによって有効性の評価も難しい．そこで，植物の栽培データといった不均衡性を持つデータでの回帰問題に対し，適切な機械学習を可能とするリサンプリング手法の検討を進めた．

3.2 Random sampling

無作為にデータを選択してリサンプリングする手法で，アンダーサンプリングを行う Random Under Sampling[8]（以降，RU），オーバーサンプリングを行う Random Over Sampling[9]（以降，RO）がある．データの多数グループと少数グループは，ユーザが入力するデータの希少度と閾値によって分割されリサンプリングが適用される．単純かつ高速で実行可能なアルゴリズムだが，無作為にデータを選択するため，重要なデータの消失や意味のないデータが複製される可能性がある．偏りが大きいデータでは多量のデータがリサンプリングされるため，閾値によってはリサンプリング後のデータ数が大きく変動する場合もある．

3.3 SMOTER

SMOTER (Synthetic Minority Over-Sampling Technique for Regression)[10]は，RU や RO と同様にデータの希少度と閾値からデータを多数グループと少数グループに分割する．多数グループに対しては RU を適用し，少数グループに対してはクラス分類問題のオーバーサンプリング手法の SMOTE[11]を回帰問題向けに改良した手法を適用する．SMOTE は k-近傍法を用いて同一クラス内のデータを 2 点選択し，点と点を結ぶ線形上に新しいデータを内挿して生成を行う．SMOTER では，クラス分類問題と異なり目的変

© 2021 Information Processing Society of Japan

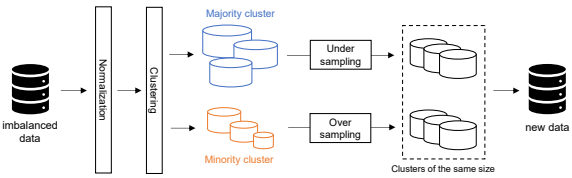

図 1　CREAMER 概要図

アルゴリズム 1　疑似コード

Algorithm 1: CREAMER

Input:

$D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ // Dataset
k // Number of clusters
ts // Target scale

Output:

D_{new} // Resampled dataset

procedure:

$D_{cls} \leftarrow normalize(D)$
$y_{ts} \leftarrow (y_i \mid y_i \in D_{cls}) \times ts$
$D_{cls} \leftarrow (x_i \mid x_i \in D_{cls}) \cup y_{ts}$
$Cls \leftarrow k - means(D_{cls})$
$msize \leftarrow mean(size(Cls))$
$Cls_{majority} \leftarrow (C_i \mid size(C_i) > msize, C_i \in Cls)$
$Cls_{minority} \leftarrow (C_i \mid size(C_i) < msize, C_i \in Cls)$
foreach Cc **in** $Cls_{majority}$
  $d_{new} \leftarrow undersampling(Cc, msize)$
  $d_{under} \leftarrow d_{new} \cup d_{under}$
end foreach
foreach Cr **in** $Cls_{minority}$
  $d_{new} \leftarrow oversampling(Cc, msize)$
  $d_{over} \leftarrow d_{new} \cup d_{over}$
end foreach
$D_{new} \leftarrow d_{under} \cup d_{over}$
return D_{new}

数が連続変数で同一でないことを考慮して，選択した 2 点と生成したデータ間の距離を重みとして，選択した 2 点の目的変数の加重平均を新しいデータの目的変数として採用するという処理を行なっている．SMOTER はオーバーサンプリングとアンダーサンプリングの両方を行うため，データ数の変動を抑えたまま不均衡性を改善できる．また，同一データを生成しないため，過学習の抑制も期待できる．

3.4 WERCS

WERCS (Weighted Relevance-based Combination Strategy) [9]は，データの希少度に応じてアンダーサンプリングとオーバーサンプリングを行う手法である．オーバーサンプリングでは与えられた希少度を確率分布とみなし，確率的にデータを選択し複製する．アンダーサンプリングでは希少度分布を反転した確率分布に基づいてデータを選択し除去する．つまり，データの全範囲に対してアンダーサンプリングとオーバーサンプリングを適用する．閾値の入力が不要であることに加え，閾値付近の類似したデータに異なるリサンプリングを適用されることがない．

3.5 データ希少度

3.2 節から 3.4 節で挙げた手法は，いずれもデータの希少度の設定が必要になる．本来であればデータのドメイン知識に基づいた希少度の決定が理想であるが困難であるため，いずれの手法においても Ribeiro が提案した関連性関数[12]によって設定している．この関連性関数は，データ希少度は目的変数の分布に反比例する，という仮定に基づいたもので，例えば目的変数が標準正規分布の場合，0 に近いほど希少度は小さく絶対値が大きくなるほど希少度は大きく設定される．

4. CREAMER

4.1 要件

本研究では栽培データの持つ分布不均衡性を解消し，適切な機械学習による推定精度の向上を目指し，新たなリサンプリング手法を提案する．ここで，既存手法の課題点も含め，次の 2 つの項目について考慮すべきと考えた．

- 栽培データは収集に時間を要しデータ数が少ない傾向にあるため，データ数の大幅な増減は不自然なデータセットを生成しやすい．従って，アンダーサンプリングとオーバーサンプリングを併用し，全体的に分布を変換すべきである．
- 植物は周囲環境に応答し，光合成や蒸散といった生理活動を行うがその関係性は単純ではない[13][14]．つまり，植物生理状態と環境条件は 1 対多の関係であり，データ希少度は目的変数のみから決定できない．しかし，説明変数を含むすべての変数から希少度を設定するのは高度な知識や分析を必要とする．そこで，収集した全変数から自動的に希少度が決定されるとよい．

4.2 提案手法の概要

栽培データの不均衡性解消を目的としたリサンプリング手法として，CREAMER (Clustering-based REsAmpling MEthod for Regression) を提案する．図 1 に CREAMER の概要図，アルゴリズム 1 に疑似コードを示す．CREAMER は既存手法と異なり，データを入力すると希少度の決定が自動的に行われ，その希少度に従ってアンダーサンプリングまたはオーバーサンプリングが適用される．また，希少度決定部では，クラスタリングアルゴリズムを用いて全ての変数の関係性を考慮することで，植物生理状態だけでなく環境条件との組みあわせを考慮した希少度が算出される．以降，各ステップの詳細を述べる．

4.3 クラスタリング前処理

クラスタリングの前処理としてデータの正規化と目的変数のスケール変換を行う．クラスタリングはデータ間距離を用いるため，単位が大きく異なる変数を含むデータで

© 2021 Information Processing Society of Japan

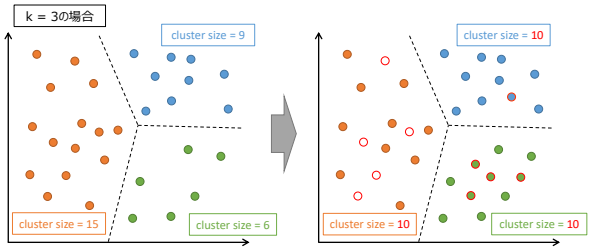

図 2　リサンプリングの例

は適切なクラスタリング結果を得られない．従って，全ての変数を$[0,1]$にスケール変換を行い，クラスタリング時の寄与率を統一する．しかし，目的変数はモデル学習時の損失に直接影響するため，目的変数のみ$ts(\geq 1)$倍の重みを追加し，$[0,ts]$にスケール変換を行う．重み付けにより目的変数は説明変数よりもクラスタ境界の決定に寄与しやすくなり，より機械学習に適した希少度を設定可能にする．

4.4　データ希少度決定

4.3 項で作成したデータを用いてクラスタリングを行い，希少度を決定する．k-means[15]によりデータをk個のクラスタに分割する．各クラスタ内のデータ数（以降，クラスタサイズ）を希少度とし，平均値より大きいクラスタサイズを持つクラスタ群を多数グループ，平均値より小さいクラスタサイズを持つクラスタ群を少数グループに分割する．クラスタリングとクラスタサイズを利用することで，全ての変数での類似度を考慮した希少度の決定を実現する．

4.5　リサンプリング

多数グループに対してはアンダーサンプリング，少数グループに対してはオーバーサンプリングを行う．この時，クラスタサイズの平均値と同じサイズになるようにリサンプリングすることで，リサンプリング前後でのデータ数の変動を抑えることができる．図 2 に$k = 3$のときのリサンプリング例を示す．アンダーサンプリングとオーバーサンプリングには任意のアルゴリズムを適用でき，既存手法と組み合わせることも可能である．以上の処理により，各環境条件をクラスタとして保持したまま全体のデータ不均衡性の解消が期待できる．

5.　検証実験

植物の光合成速度と蒸発散速度の推定についてリサンプリングの有効性と CREAMER の有効性を確認するため，実際の栽培データを用いて検証実験を行う．

5.1　データセット

イチゴの品種‘きらぴ香’を対象とし，2019 年 11 月 1 日から 2020 年 6 月 11 日にかけて静岡県農業技術産学官連携

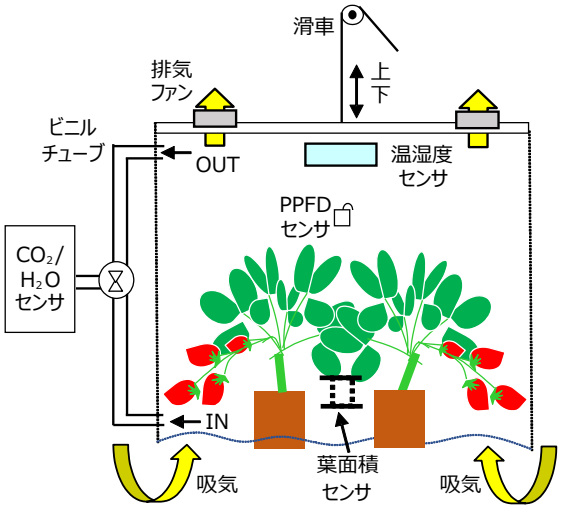

図 3　センサ設置位置

表 1　データセット詳細

説明変数	環境データ	温度[℃]，湿度[%]，飽差[kPa]，CO_2[ppm]，PPFD[$\mu mol\ m^{-2}\ s^{-1}$]
	植物データ	NIR/VR（葉面積）
	時系列特徴	日の出からの経過時間[分]
目的変数		光合成速度[$\mu mol\ p^{-1}\ s^{-1}$]，蒸発散速度[$g\ p^{-1}\ min^{-1}$]
収集データ数		16334
データ周期間隔		10分（目的変数），1分（説明変数）
データ収集期間		2019/11/01 ~ 2020/06/11
データ収集時間		0:00 ~ 24:00

研究開発センター（静岡県沼津市西野）に設置された温室内でデータセットの収集が行われた．光合成速度と蒸発散速度の計測はチャンバー法[16]を採用し，データ収集はチャンバー内にセンサを設置して実施した（図 3）．表 1 にデータセットの詳細を示す．

説明変数は環境データ 5 次元，植物データ 1 次元，時系列特徴量 1 次元の計 7 次元の変数を持つ．葉面積の計測では，葉の光の吸収率が波長によって異なる性質を利用するイチゴ用葉面積モニタリングセンサ[17]を使用している．また，本実験で用いるリサンプリング手法は時系列性を考慮しておらず，RNN や LSTM[18]のような時系列データに有効なモデルを使用できないため時系列特徴量を追加した．具体的には，予測対象と関連のあるイベントを起点として算出した時系列特徴が有効であるという報告[19]を参考に，植物の生理活動と関係のある日の出からの経過時間を説明変数に追加した．

光合成速度と蒸発散速度は，それぞれ式(1)，式(2)に従って算出した．ここで，F_rは空気の流量$[m^3 min^{-1}]$，ΔCO_2はCO_2濃度差$[ppm]$，T_{in}はチャンバー内気温[℃]，ΔH_2OはH_2O

© 2021 Information Processing Society of Japan

濃度差[gm^{-3}]を示す．

$$CER = \frac{(F_r \cdot \Delta CO_2)/(0.0224 \cdot (273 + T_{in})/273)}{60 \cdot 6} \quad (1)$$

$$ETR = F_r \cdot \Delta H_2O/6 \quad (2)$$

目的変数は 10 分周期で算出するため，説明変数は 10 分間の平均値に変換して使用した．機械学習を用いるにあたり，日毎のデータは時間的依存性が大きくないと考え，訓練データ，検証データ，テストデータは各月ごとに日単位で 6:2:2 に分割することとした．

5.2 実験方法

栽培データに対するリサンプリングの有効性及び CREAMER の有効性を検証するために，前処理を行わない（以降，RAW）データと既存手法の RU，RO，SMOTER，WERCS を適用したデータで，機械学習の推定精度について比較を行った．表 2 に各手法におけるパラメータの組み合わせを示す．ここで，SMT，WRS，CRM はそれぞれ SMOTER, WERCS, CREAMER を指す．CREAMER のリサンプリング部は任意な手法が選択可能なため，RU と RO を組み込んだ CRM_RU_RO と，RU と SMOTER の内挿処理部を組み込んだ CRM_RU_SMT で試行した．既存手法で入力するデータ希少度のパラメータ'relevance'には，3.5 説で述べた関連性関数[12]を利用した．図 4 に光合成速度と蒸発散速度の分布と希少度を示す．どちらの指標とも 0.0 付近にデータが集中し，値が大きくなるほど顕著にデータ数が減少している．両指標とも 0 未満の数値はセンサのノイズであることが多く，推定において重要な値でないため，0 未満のデータ希少度を 0 に置換した．

表 2　パラメータ一覧

手法	パラメータ
RAW	(None)
RU	relevance=[rarity], relevance_threshold=[0.4, 0.5, 0.6], under=['balance']
RO	relevance=[rarity], relevance_threshold=[0.4, 0.5, 0.6], over=['balance']
SMT	relevance=[rarity], relevance_threshold=[0.4, 0.5, 0.6], over=['balance']
WRS	relevance=[rarity], under=[0.5, 0.8], over=[0.5, 0.8]
CRM_RU_RO	k=[3, 10, 20], ts=[3.0, 4.0]
CRM_RU_SMT	k=[3, 10, 20], ts=[3.0, 4.0]

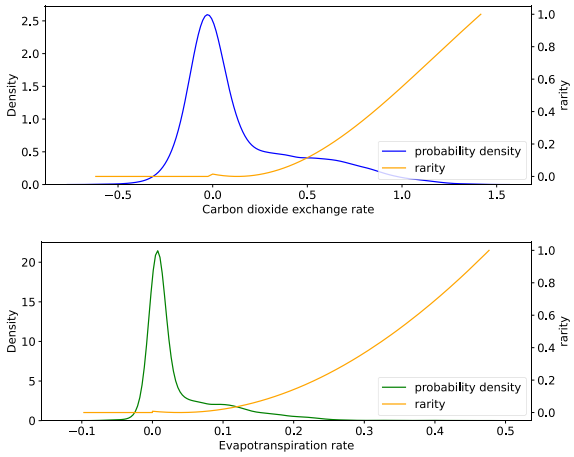

図 4　目的変数の分布と希少度
（上：光合成速度，下：蒸発散速度）

機械学習は MLP を用いて行い，隠れ層の数を 1，3，5 層（各ノード数は 64），学習率を 0.1, 0.01 で変化させて試行を行った．最終的なモデル選択は検証データに対する平均平方二乗誤差が最も低くなるリサンプリング手法のパラメータと MLP のパラメータの組み合わせを採用した．各モデルの推定値について平均平方誤差（MAE）と平均平方二乗誤差（RMSE）の誤差指標で精度比較を行った．

5.3 光合成速度の推定

図 5 に光合成速度推定におけるリサンプリング手法ごとのテストデータに対する推定誤差を示す．リサンプリングを行わない RAW では，MAE が 0.128，RMSE が 0.032 であり，既存手法と比較すると RAW の方が，誤差が小さいという結果になった．一方で，CRM_RU_RO では MAE が 0.124, RMSE が 0.032, CRM_RU_SMT では MAE が 0.120, RMSE が 0.029 であり，僅かながら CREAMER を用いることで，最大で MAE を 6.25%，RMSE を 9.38%削減できた．

以上の結果より，光合成速度の推定ではリサンプリングによる恩恵は少ないことが分かったが，CREAMER を適用することで精度の向上が確認できた．特に，SMT と CRM_RU_SMT はリサンプリング処理が同じでありながら推定誤差に差があることから，クラスタリングによる説明変数も考慮したデータ希少度の設定が影響したといえる．

5.4 蒸発散速度の推定

図 6 に蒸発散速度推定におけるリサンプリング手法ごとのテストデータに対する推定誤差を示す. RAW と比較すると, RO 以外ではリサンプリングを適用することで, 推定誤差が削減された．各リサンプリング手法の推定誤差を比較すると, CRM_RU_RO は MAE が 0.0211, RMSE が 0.0014 と最も誤差が小さく，次点で CRM_RU_SMT の誤差が小さいという結果になった．また，CRM_RU_RO を RAW と比較すると MAE で 53.11%，RMSE で 40.74%の誤差を削減できた．

以上の結果より，光合成速度とは異なり蒸発散速度の推定では，リサンプリングの適用が非常に有効であることが確認できた．これは，図 4 より蒸発散速度の分布は，光合成速度の分布より偏りが大きく，推定精度の低下に対する

© 2021 Information Processing Society of Japan

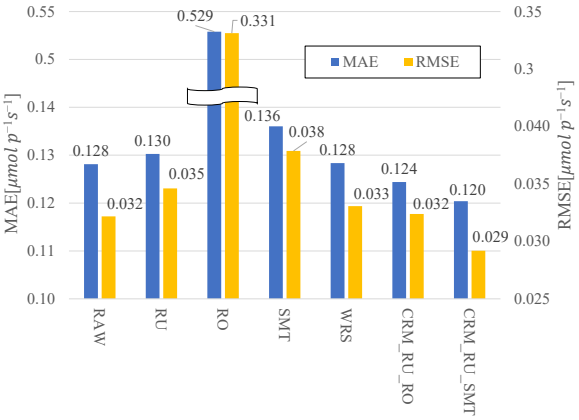

図 5　光合成速度推定結果

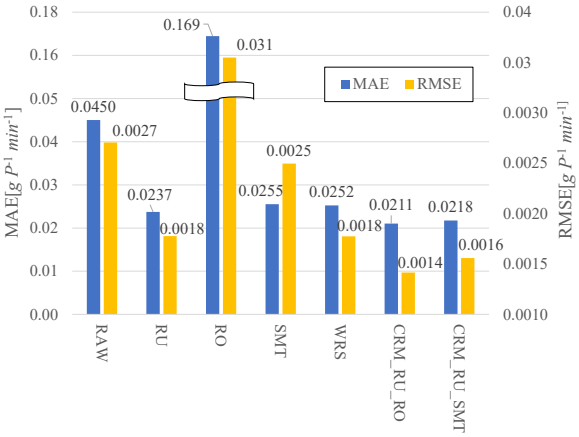

図 6　蒸発散速度推定結果

不均衡性の影響度合いも大きくなるため，光合成速度の推定よりもリサンプリングの効果を得られたと考える．また，CREAMER については，試行した 2 手法とも既存手法以上の推定精度の向上を確認した．

5.5 分布変化の分析

図 7 に各リサンプリング手法適用後の目的変数の分布，表 3 に目的変数の尖度と歪度を示す．表 3 の赤文字箇所は，連続一様分布の歪度 0.0，尖度-1.2 に近い上位 3 つを示す．図 7 からリサンプリングにより目的変数の分布が変動し，分布の不均衡性が低減していることが分かる．いずれの手法も両方の目的変数について 0.0 付近のデータ割合が削減しているが，分布の形状に違いがあるように見える．RU, RO, SMOTER は，光合成速度では 1.0 付近，蒸発散速度では 0.4 付近のデータ割合が増加し，二峰性分布になっている．多数グループと少数グループを希少度の閾値で分割しており，閾値の前後で異なるリサンプリングが適用されることが原因と考える．WERCS は，光合成速度では歪度，尖度が非常に一様分布と近く，図 7 からも最も平坦に近い分布であると定性的に分かるが，蒸発散速度に対しては 0.0 付近のデータが一定割合残っている．蒸発散速度の

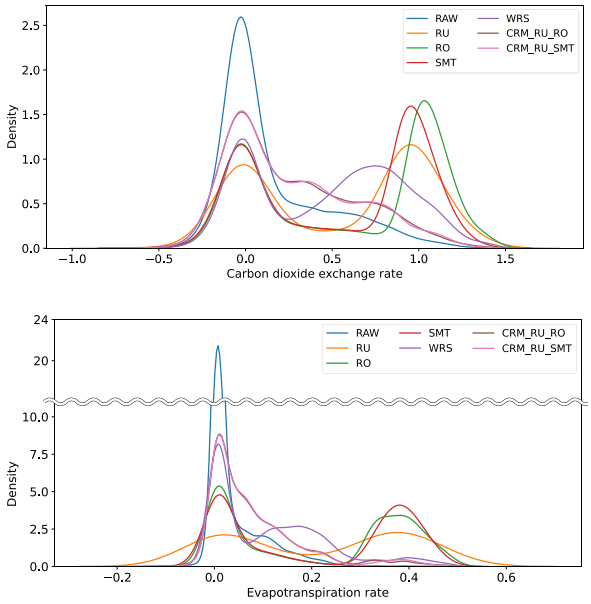

図 7　各リサンプリングでのデータ分布
（上：光合成速度，下：蒸発散速度）

表 3　各リサンプリングでの歪度と尖度

	光合成速度		蒸発散速度	
	歪度	尖度	歪度	尖度
RAW	1.240	0.923	2.224	6.349
RU	*-0.205*	*-1.585*	*-0.074*	*-1.769*
RO	*-0.248*	-1.624	*-0.014*	*-1.775*
SMT	*-0.248*	*-1.604*	*-0.049*	*-1.817*
WRS	*0.006*	*-1.224*	1.029	0.430
CRM_RU_RO	0.646	-0.452	1.739	3.183
CRM_RU_SMT	0.644	-0.477	1.758	3.177

分布は光合成速度より不均衡度合いが大きく，WERCS は，全体のデータ数に対してリサンプリングの対象データ数を決定するため，十分にリサンプリングができなかった可能性がある．CREAMER は，RAW の尖度のみを低くしたような分布となり，目的変数の不均衡性解消はしきれていないといえる．要因として，CREAMER では，希少度の設定に説明変数も加味していることが影響している可能性がある．ここで，前節の推定精度比較結果と表 3 を照らし合わせると，目的変数が理想の分布に近い分布であっても必ずしも推定精度が向上している訳ではない．つまり，機械学習にとって目的変数の不均衡性の解消だけが，最善策ではないと考えることもできる．以上より，CREAMER は全ての変数を重畳する希少度決定手法によって，不均衡性の解消だけでなく，機械学習に適したリサンプリングができると考える．

図 8，図 9 に各リサンプリング手法適用後の説明変数の分布を示す．可視化のための次元削減として UMAP[20]を使用した．RAW データの説明変数に対して学習を行い，7

© 2021 Information Processing Society of Japan

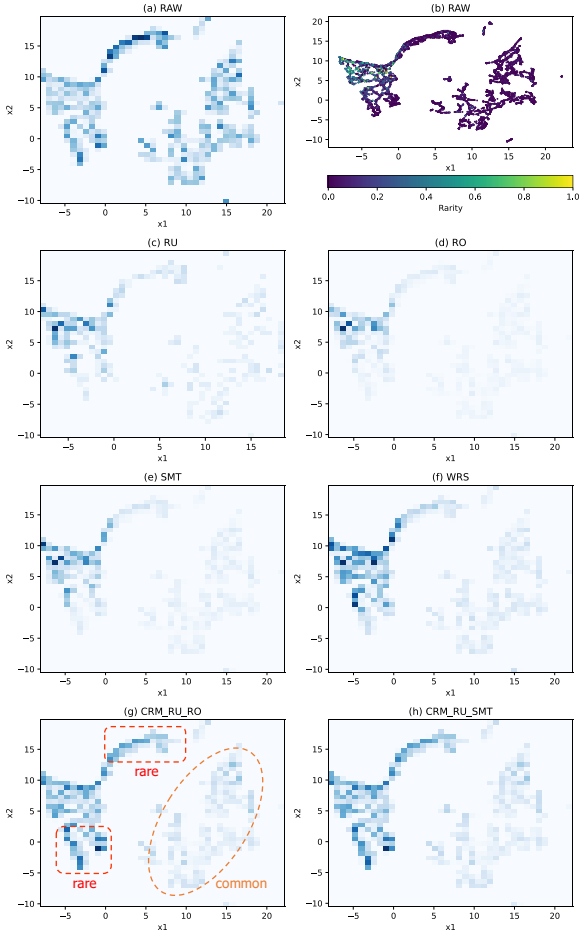

図 8　説明変数の分布（光合成速度）

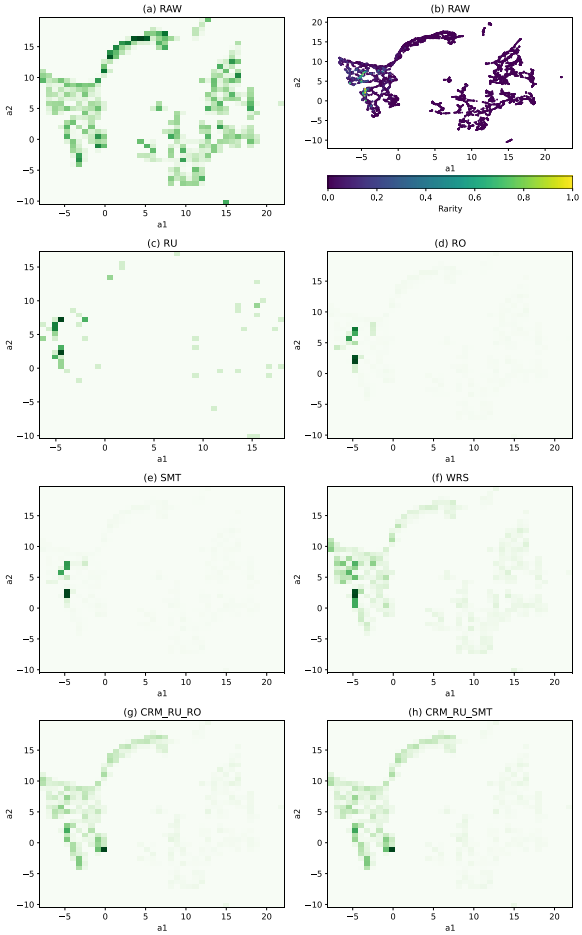

図 9　説明変数の分布（蒸発散速度）

次元から 2 次元に次元削減を行い 2 変量ヒストグラムで表現した．また，図中(b)は，RAW データの目的変数の希少度との関係を濃淡散布図で表した．図 8，図 9 から，既存手法での説明変数の分布は，目的変数の希少度が高い部分にデータが集中した分布と類似していることが分かる．それに対して CREAMER は，均等にデータを分散するような分布を形成しており，類似した目的変数を持つデータに対しても異なる扱いをしていることが分かる．例えば，図 8 中(g)の $x2$ が 15 付近のデータ群や，$x1$ と$x2$ がともに 0 付近のデータ群は，$x1$が 5 以下のデータ群と同じような目的変数を持っているが，リサンプリング後のデータ密度が異なっている．さらに，説明変数が密な部分に対して，目的変数の希少度が高い部分のみのデータを残していることから，クラスタリング時の目的変数の重みづけが有効に働いているといえる．以上の結果から，CREAMER ではクラスタリングによる希少度決定手法により，説明変数の偏りを防げることを定性的に確認できた．説明変数のばらつきを保ちながら目的変数の不均衡性を解消したことで，モデル学習時の偏りが抑制され，検証実験において既存手法を上回る精度を得ることができたと考える．

6. おわりに

本研究では，植物の光合成速度及び蒸発散速度の推定において，目的変数と説明変数の両方を加味したデータ希少度を自動的に決定し，機械学習に適した不均衡性分布の変換処理を行う CREAMER を提案した．CREAMER の有効性の確認を目的とし，イチゴの栽培データによる検証実験を行なった．光合成速度の推定ではリサンプリングを行わない場合と比較して既存手法は逆に誤差が大きくなったが，唯一 CREAMER は誤差を削減することに成功した．蒸発散速度の推定では多くの手法で誤差の削減に成功したが，CREAMER は最も誤差を削減し，リサンプリングを行わない場合と比較して約半分の誤差での推定を可能にした．

今後の方針として，時系列を考慮したリサンプリングの検討を行う．センサデータでは植物生理状態の表現に限界があり，植物体の成長や概日リズムによる生理活動の変化の補足は困難である．そういったトレンド成分や季節成分といった時系列関係を表現するため，時系列情報の保持が可能なリサンプリング手法を考案する．さらに，CREAMER の汎用性を確かめるため，農業関連に限らず様々なデータセットへの適応性を検証する．株価，電気需要，通信トラ

© 2021 Information Processing Society of Japan

フィックなど分布が不均衡かつ少数グループの推定・予測が重要な問題は少なくなく，活用面は多くあると考える．多種のデータセットで検証を行い，アルゴリズムの改善を繰り返して，汎用的に使用可能な手法の実現を目指す．

謝辞

本研究の一部は，JST 創発的研究支援事業（JPMJFR201B）の支援を受けたものである．また，データセットの提供及び農学知見を支援いただいた静岡県農林技術研究所に感謝の意を表する．

参考文献

[1] 農林水産省: スマート農業の展開について，農林水産省（オンライン），入手先 <https://www.maff.go.jp/j/kanbo/smart/pdf/smart_agri_tenkai.pdf> (参照 2021-04-27).
[2] 日本農業気象学会: 新編 農業気象学用語解説集 －生物生産と環境の科学－，日本農業気象学会，pp.203-204 (1997).
[3] 渡部一郎: 農業環境実験法＜農業気象学・農業環境工学＞，pp.99-112，サイエンスハウス (1987).
[4] 日本農業気象学会: 新編 農業気象学用語解説集 －生物生産と環境の科学－，pp.191，日本農業気象学会 (1997).
[5] 渡部一郎: 農業環境実験法＜農業気象学・農業環境工学＞，pp.113-120，サイエンスハウス (1987).
[6] 亀山幸司，宮本輝仁，岩田幸良: 土壌水分計を用いた圃場内の水分分布測定に基づく代表地点の選出，農業農村工学会論文集，Vol.87，No.2，pp.Ⅱ_113-Ⅱ_121 (2019).
[7] Krawczyk, B.: Learning from imbalanced data: open challenges and future directions, Progress in artificial intelligence, Vol.5, No.4, pp.221-232 (2016).
[8] Torgo, L., Branco, Paula. and Ribeiro, RP.: Resampling strategies for regression, Expert systems, Vol.32, No.3, pp.465-476 (2015).
[9] Branco, Paula., Torgo, L. and Ribeiro, RP.: Pre-processing approaches for imbalanced distributions in regression, Vol.343, pp.76-99 (2019).
[10] Torgo, L., Branco, Paula., Pfahringer, B., et al.: Smote for regression, Portuguese conference on artificial intelligence, pp.378-389 (2013).
[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., et al.: SMOTE: Synthetic Minority Over-sampling TEchnique, Journal of artificial intelligence research, No.16, pp.341-378 (2002).
[12] Torgo, L., Ribeiro, RP.: Utility-Based Regression, European conference on principles of data mining and knowledge discovery, pp.597-604 (2007).
[13] 和田義春，添野隆史，稲葉幸雄: 促成，半促成栽培におけるイチゴ品種‘とちおとめ’の 高 CO2 濃度下の葉光合成速度促進に及ぼす光と温度の影響,Vol.79, No.2, pp.192-197 (2010).
[14] 稲葉幸雄: “いちご「とちおとめ」の蒸散量”, Vol.20, pp.45-46 (2001).
[15] MacQueen, J.: Some methods for classification and analysis of multivariate observations, Proc. Fifth berkeley symposium on mathematical statistics and probability, pp.281-297 (1967).
[16] 高山弘太郎，下元耕太，高橋憲子ほか: 太陽光利用型植物工場におけるトマト個体光合成蒸散リアルタイムモニタリング，日本生物環境工学会, pp. 216-217 (2012).
[17] 大石直記，貫井秀樹，佐藤陽介: 人工光源下の植物成育評価を目的としたコンパクト散乱光センサの開発，日本生物環境工学会, pp. 18-19 (2018).
[18] Hochreiter, S., Schmidhuber, J.: Long short-term memory, Neural computation, Vol.9, No.8. pp.1735-1780 (1997).
[19] 若森和昌，柴田瞬，澤村武ほか: ニューラルネットワークを用いた植物蒸発散量推定における時系列特徴重畳手法,電子情報通信学会技術研究報告 信学技報, Vol.117, No.442, pp.115-120 (2018).
[20] Leland, M., John, H., and Melville J.: Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

© 2021 Information Processing Society of Japan