水稲の収量･品質予測手法としてのサポートベクターマシーンの評価

誌名	九州大学大学院農学研究院学芸雑誌
ISSN	13470159
著者	髙城, 大地 猿田, 恵輔 平井, 康丸 井上, 英二 岡安, 崇史 光岡, 宗司
巻/号	67巻2号
掲載ページ	p. 47-52
発行年月	2012年9月

農林水産省 農林水産技術会議事務局筑波産学連携支援センター
Tsukuba Business-Academia Cooperation Support Center, Agriculture, Forestry and Fisheries Research Council Secretariat

AgriKnowledge

九大農学芸誌 (Sci. Bull. Fac. Agr., Kyushu Univ.)
第67巻 第2号 47-52 (2012)

水稲の収量・品質予測手法としてのサポートベクターマシーンの評価

髙城大地[1]・猿田恵輔[2]・平井康丸[3]*
井上英二[3]・岡安崇史[3]・光岡宗司[3]
九州大学大学院農学研究院環境農学部門生産環境科学講座生物生産工学研究分野

(2012年4月23日受付, 2012年5月10日受理)

Evaluation of Support Vector Machine as an Analytical Method for Building Prediction Models of Rice Yield and Quality

Daichi TAKAJYO[1], Keisuke SARUTA[2], Yasumaru HIRAI[3]*, Eiji INOUE[3], Takashi OKAYASU[3] and Muneshi MITSUOKA[3]

Laboratory of Bioproduction Engineering, Division of Bioproduction Environmental Sciences, Department of Agro-environmental Sciences, Faculty of Agriculture, Kyushu University, Fukuoka 812-8581, Japan

緒　　言

我が国の水稲生産は，近年の夏期高温等の影響により，九州地域を中心として一等米比率の低下が著しい(農林水産省, 2011)．また，消費者の高品質志向，国際競争への対応が求められるが，依然として，農家間や年次間で収量・品質が不安定であり，その高位安定化が重要な課題となっている．収量・品質の不安定の原因は，農業が天候を始めとする多くの要因の影響を受ける側面を持つことから，作業の意思決定が農家の経験や勘に依存する部分が大きいためと考える．この経験や勘に依存した生産方法に起因する不安定を低減するためには，生産管理，生産環境，生育状態，収量・品質等の数値情報に基づく生産技術の確立が必要であり，その支援技術の開発が求められる．

以上の背景から，本研究では，水稲生産における農作業意思決定支援に活用しうる収量・品質予測モデルの開発を行っている．従来，水稲の収量・品質予測に関しては，気象と水稲の生育を関連づけたモデルSimulation Model for Rice-Weather Relations (SIMRIW) により収量の予測のみが可能になっている．このモデルは機構的モデルであるため，発育速度については栽培試験を通した品種パラメータの推定が必要である．また，乾物重生産や収量に関するパラメータは，ジャポニカ米とインディカ米の区別しかない等 (田中ら, 2011)，広域レベルで水稲の栽培可能性や最適品種を判断する用途への適性が高い．一方，本研究が開発を進めるモデルは，近年の各種情報収集機器の著しい発展を踏まえた，生産現場で収集されるデータ (事例) に基づく学習型のモデルであり，水田一枚レベルの予測を対象

[1] 九州大学大学院生物資源環境科学府環境農学専攻生産環境科学教育コース生産環境情報学分野
[2] ヤンマー農機株式会社
[3] 九州大学大学院農学研究院環境農学部門生産環境科学講座生物生産工学研究分野
[1] Laboratory of Bioproduction and Environment Information Sciences, Course of Bioproduction Environmental Sciences, Department of Agro-environmental Sciences, Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University
[2] YANMAR AGRICULTURAL MACHINERY MANUFACTURING CO., LTD.
[3] Laboratory of Bioproduction Engineering, Division of Bioproduction Environmental Sciences, Department of Agro-environmental Sciences, Faculty of Agriculture, Kyushu University
* Corresponding author (E-mail: hirai@bpes.kyushu-u.ac.jp)

とする．本モデルの開発により，生産環境，生産管理，生育状態と収量・品質の関係が構築され，肥培管理や水管理等の作業に関する意思決定支援が可能になると考える．

予測モデルに求められる要件としては，①生産現場で収集される大量・多様な情報を使って説明変数と収量・品質の複雑な関係を構築できる，②各水田において収集される代表データの揺らぎ（ばらつき）に対して頑強である，③生産者が予測結果を容易に理解・解釈できることが挙げられる．

本研究では，上記3つの要件を満たす収量・品質予測モデルを作成する際の基盤解析手法として，学習型パターン認識手法であるサポートベクターマシン（以下，SVM）の評価を行った．まず，Directed Acyclic Graph（以下，DAG）により多クラス判別を可能にしたDAGSVMを用いて，出穂期以降の稲体の状態量（窒素栄養，収量キャパシティ）および頴花分化終期（出穂14日前）以降の気象環境（気温，日射）から，精玄米収量および玄米タンパク質含有率のパターンを予測するモデルを作成した．さらに，判別率およびソフトマージンによる汎化能力の調整機能の評価を通して，SVMの予測モデル作成手法としての有効性を検討した．

本研究の遂行に当たり九州大学大学院農学研究院植物栄養学分野の山川武夫准教授には，玄米のタンパク質含有率の分析に関してご指導頂きました．ここに記して感謝の意を表します．

材料及び方法

1．2クラスサポートベクターマシーンの理論

サポートベクターマシーン（以下，SVM）は学習型の線形2群判別器である．ただし，学習データを写像した高次元特徴空間上で2群の分離超平面を求めることにより，元のデータ空間における非線形判別器を構成することができる．これにより，複雑な境界を有するパターンの分類を可能にしている．さらに，マージン（分離超平面と最も近いデータとの距離）を最大化する超平面を求めることで，高い汎化能力を実現する．また，線形分離出来ないデータについては，ソフトマージンと呼ばれる機能により，学習に及ぼす影響を増減させることで汎化能力の調整が可能である．ここで，2つのクラスG_1，G_2の判別器を作成するSVMの学習は，数理計画法の手法を用いると，以下の最適化問題に帰着される．

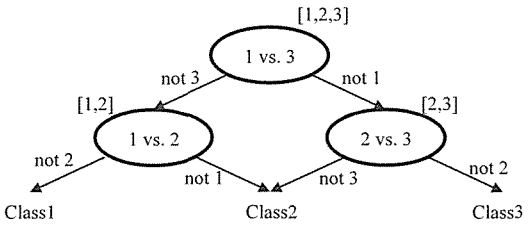
$$\max_{\alpha}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right\}$$
$$s.t.\quad \sum_{i=1}^{n}\alpha_i y_i=0, \qquad (1)$$
$$0\le\alpha_i\le C,$$
$$i=1,2,\ldots,n$$

ここで，Cは線形分離されないデータに対するペナルティを調整するパラメータ，i，jは学習データの番号を示す添え字，kは学習データの高次元空間への写像と学習上必要な内積計算を同時に行うカーネル関数，nは学習データ数，$\boldsymbol{x}_i$は説明変数データベクトル，y_iは1もしくは−1の値をとり，それぞれ学習データ$\boldsymbol{x}_i$がG_1，G_2に属することを示す．α_iはラグランジュ乗数であり，各学習データに対して1つずつ割り当てられる．また，(1) の最適化問題に対する解として求まるα_iは次の様な意味を持つ．a) $\alpha_i=0$のとき，$\boldsymbol{x}_i$は境界から内部にあり分離超平面の構成に寄与しない．b) $0<\alpha_i<C$のとき$\boldsymbol{x}_i$は境界と近接し超平面構成に寄与している．なお，このときの$\boldsymbol{x}_i$はサポートベクターと呼ばれる．c) $\alpha_i=C$のとき，$\boldsymbol{x}_i$は線形分離されていない．さらに，求めたα_iから判別に用いる閾値bを計算すると，SVMによる判別方法は次に与えられる．

$$x\begin{cases}\in G_1 & \left(\sum_{j=1}^{n}y_j\alpha_j k(\boldsymbol{x}_j,\boldsymbol{x})+b\ge 0\right)\\ \in G_2 & \left(\sum_{j=1}^{n}y_j\alpha_j k(\boldsymbol{x}_j,\boldsymbol{x})+b<0\right)\end{cases} \qquad (2)$$

2．Directed Acyclic Graphによる多クラス化

Directed Acyclic Graph（以下，DAG）を用いて，2クラスの判別器であるSVMを多クラスの判別器（DAGSVM）に拡張した（Platt *et al.*, 2000）．DAGSVMでは，全てのクラスの組み合わせでSVMを構築した後，図1に示す判別方式に従い，各SVMを結合する．すなわち，全てのクラス番号を要素とするリストを作成し，クラスi vs. jの判別で，クラスjでないと判別された場合リストからjを除外する．この手順を繰り返し，最後に残る番号が帰属クラスとなる．DAGSVMの利点は計算の迅速性にある（阿部，2008）．これは各SVMが全学習データの一部で学習を行い，判別に用いるSVMも限られ，最終的な演算回数が少ないためである．



〔注〕図中の数字はクラス番号を，大括弧内はクラス番号のリストを示す．

図1　DAGによる分類方法

3．解析プログラム

学習のアルゴリズムには，Sequential Minimal Optimization（以下，SMO）（Platt, 1999）を用いた．本手法では2つのラグランジュ乗数を適宜選択して生成される部分問題を繰り返し解き，最適解を求める．各部分問題は（1）式の制約条件を利用すると解析的に解が求められ，高速な学習が可能である．解析プログラムはVisual Basic 2010により作成した．

4．解析に使用したデータの収集方法

解析に使用したデータは，福岡県八女市星野村（2009年），九州大学農学部附属農場（2009年，2010年），糸島市および福岡市（2010年，2011年）において実施した延べ47枚の水田調査（品種ヒノヒカリ）により得られたものである．水稲生育期間中の気温および日射量は，気象観測装置（HOBO Weather Stations, ONSET COMPUTER）により計測した．生育指標は，分げつ盛期，最高分げつ期，幼穂形成期，出穂期，登熟期の各生育期に，草丈については，4～10株，茎数およびSPAD値については10～20株を各調査水田で計測し，平均値を求めた．SPAD値は，中鉢（1986）の報告を参考にして，株中の草丈が最長となる茎の第2展開葉（登熟期のみ止葉）の中央部において，中肋を挟む左右各2箇所（合計4箇所）を葉緑素計（SPAD－502，コニカミノルタ）により計測し，平均値を求めた．収量構成要素および玄米タンパク質含有率については，各調査水田から10株を手刈りして分析試料とした．試料は通風乾燥後，各株の穂数，10株の合計籾数を計測した．さらに，インペラ籾すり機（FC2K，大竹製作所）により籾すりした玄米を，1.85mmの縦目篩にて選別し，篩上の精玄米数を計測した．籾数，精玄米数の計測には，粒数検知器（WAVER IC－0，アイデックス）を用いた．続いて，精玄米重を秤量し，試料の含水率を135°C24時間法（農業機械学会，1996）により測定した．以上の測定結果から，各試験区の穂数，籾数，登熟歩合，千粒重の収量構成要素を求めた．精玄米収量（収量）については，2009年および2010年については，収量構成要素の値から算定した．2011年については，60株の坪刈り調査により求めた．玄米の全窒素については，精玄米を48時間70～80°Cで通風乾燥後，サイクロンサンプルミル（CSM-FI，UDY CORPORATION）により粉砕した試料を，硫酸－過酸化水素分解法（大山ら，1991）により1試料につき3連制で分解した．続いて，インドフェノール法（Cataldo *et al.*, 1974）により比色定量後，平均値を算定した．ここで，吸光度の測定には，紫外可視分光光度計（V-630，日本分光）を用いた．タンパク質含有率は窒素含有率にタンパク質換算係数5.95を乗じることにより求めた．なお，千粒重，収量，タンパク質含有率は，湿量基準含水率（wet basis）で15%の値に換算した．

5．予測モデル作成用の学習データ

予測対象である収量とタンパクについては，データの値域を3等分し，値の小さい順にLow，Middle，Highの3クラス（パターン）に分類した（表1）．また，この数値データの離散化は，SVMを用いた予測モデルの作成に必要な前処理であるとともに，モデルの頑強性，予測結果の解釈の容易性に寄与するものである．パターンを予測するための説明変数は，既往の作物学の研究成果に基づき，表2に示す要素からなる7次元のベクトルで構成した．表2中の出穂を基準とし－14～－7日間および－7～＋7日の日平均気温は，それぞれ，籾殻の大きさおよび不稔に影響を及ぼす環境変量である（佐藤ら，1973）．出穂後5～30日の日平均気温および日平均日射量はデンプン蓄積を通して登熟

表1　収量・タンパク質データの分類

収量	範囲（kg m^{-2}）	データ数
Low	312-413(356)	16
Middle	413-515(455)	26
High	515-618(553)	5
タンパク質	範囲（%）	データ数
Low	5.9-6.7(6.3)	23
Middle	6.7-7.5(7.1)	15
High	7.5-8.3(7.9)	9

〔注〕括弧内の数値はパターンの平均値を示す．

表2 説明変数ベクトルの要素

変数No.	内容
1	出穂を基準とし－14～－7日の日平均気温
2	出穂を基準とし－7～＋7日の日平均気温
3	出穂後5～30日の日平均気温
4	出穂後5～30日の日平均日射
5	出穂期のm^2当たりの稲体の窒素吸収量
6	出穂約20日後のSPAD値
7	m^2当たりの籾数

に影響を及ぼす環境因子であり，既往の報告（相見ら，1956；佐藤・稲葉，1976；若松ら，2006；田中ら，2010）を総合的に判断して影響期間を決定した．出穂期のm^2当たりの稲体の窒素吸収量および出穂約20日後のSPAD値は，収量，タンパクに影響を及ぼす稲体の生育量および窒素栄養状態の指標として選定した．ここで，稲体の窒素吸収量は，荒木ら（2006）の推定式を用いて，草丈，茎数，SPAD値および気温の実測データから求めた．また，出穂約20日後のSPAD値については，玄米タンパク質含有率と$r=0.814$の高い相関が報告されている（森ら，2010）．m^2あたり籾数は，収量キャパシティの指標として説明変数に加えた．以上の説明変数ベクトルと収量・タンパクの分類（離散化）結果を組み合わせて，収量パターン作成用，タンパクパターン作成用の2種類の学習データを作成した．

6．予測モデルの作成と評価方法

水稲生産において，収量・品質を高位安定化する管理作業の意思決定支援を行うためには，代表的な生育ステージにおいて，目標とする収量・品質に対応する生産環境，生育，農家の管理の条件を明らかにする必要がある．今回作成したモデルは，出穂期以降の稲体の状態量（窒素栄養，収量キャパシティ）および穎花分化終期（出穂14日前）以降の気象環境（気温，日射）から収量およびタンパク質のパターンを予測するものであり，農家の管理作業を含まないが，穎花分化終期以降について，収量・品質に対応する生育・環境条件を明らかにすることに資するものである．

予測モデルの作成においては，まず，各SVMのパラメータσおよびCを，leave-one-outクロスバリデーションの判別率に基づき設定した．σはSVMの構成に用いた以下に示すカーネル関数固有のパラメータである．

$$k(x_i, x_j) = \exp(-\sigma \| x_i, x_j \|) \quad (3)$$

さらに，設定条件下でDAGSVMを構築し，収量およびタンパクの予測モデルを作成した．予測モデルの精度評価は，leave-one-outクロスバリデーションの判別率により行った．

結　　果

表3にSVMおよびDAGSVMの判別率を示す．表中のLow vs. Middle，Low vs. High，Middle vs. Highは，収量もしくはタンパクの2群判別器，DAGSVMは，2群判別器を組み合わせて構成される多群判別器である．DAGSVMの判別率は，収量が85.1%，タンパク質が76.6%と比較的良好であった．また，両予測モデルにおいて，Low vs. MiddleのSVMの判別率が低かった．表4に各SVMにおける学習データ中のサポートベクトルの割合を示す．Low vs. Middleでサポートベクトルの割合が高い．

表3 SVMおよびDAGSVMの判別率（%）

	判別器	収量	タンパク質
SVM	Low vs. Middle	85.7	78.9
	Low vs. High	100.0	96.9
	Middle vs. High	96.8	87.5
DAGSVM		85.1	76.6

〔注〕SVMは2群判別器，DAGSVMはSVMから構成される多群判別器である．

表4 学習データ中のサポートベクトルの割合（%）

SVM	収量	タンパク質
Low vs. Middle	40.1	40.6
Low vs. High	25.0	19.7
Middle vs. High	23.1	36.1

考　　察

クロスバリデーションにおける誤判別率は学習データ中のサポートベクトルの割合で上限が決定される（赤穂，2008）．これは，サポートベクトルになったデータは，分離超平面近傍に存在するため，クロスバリデーション時に，学習データから除外すると，超平面の構成に影響し，誤判別される可能性があるからである．したがって，Low vs. Middleにおける低判別率（表3）は，サポートベクトルの割合の高さ（表4）に起因すると考えられる．また，収量，タンパク質の各パターン予測モデルのLow vs. Middleにおいて線形分離され

ない学習データが，それぞれ4.6%，7.5%存在した．そこで，パラメータを調整し，データを完全に線形分離できるように調整したところ，逆に判別率は低下した．よって，Low vs. Middle ではソフトマージンによる汎化能力の調整が有効に機能し，比較的高い判別率を保っていると考えられた．一方，他のSVMでは全学習データが完全に線形分離されていた．本研究を通して，作成モデルの比較的高い判別率とソフトマージンによる汎化能力の調整機能が確認された．このことから，多数の因子の影響を受ける水稲の収量・品質の予測モデル作成において，SVMは精度と汎化能力を保証し得る有効な手法であることが示唆された．

要　　約

水稲の収量・品質予測モデルを作成する際の基盤解析手法として，学習型パターン認識手法であるサポートベクターマシン（以下，SVM）の評価を行った．まず，Directed Acyclic Graph（以下，DAG）により多クラス判別を可能にしたDAGSVMを用いて，精玄米収量（以下，収量）および玄米タンパク質含有率（以下，タンパク）のパターン予測モデルを作成した．今回作成したモデルは，出穂期以降の稲体の状態量（窒素栄養，収量キャパシティ）および穎花分化終期（出穂14日前）以降の気象環境（気温，日射）の7変数から収量およびタンパクのパターンを予測するものである．モデル作成に使用したデータは，福岡県八女市星野村（2009年），九州大学農学部附属農場（2009年，2010年），糸島市・福岡市（2010年，2011年）において実施した延べ47枚の水田調査（品種ヒノヒカリ）により得られたものである．さらに，SVMの予測モデル作成手法としての有効性を，判別率およびソフトマージンによる汎化能力の調整機能の面から検討した．その結果，収量およびタンパクの判別率は，それぞれ，85.1%，76.6%と比較的良好であった．また，線形分離されない学習データに対して，ソフトマージンによる汎化能力の調整が有効に機能したことを確認した．以上の結果から，多数の因子の影響を受ける水稲の収量・品質の予測モデル作成において，SVMは精度と汎化能力を保証し得る有効な手法であることが示唆された．

キ　ー　ワ　ー　ド

米生産，サポートベクターマシーン，収量，タンパク質，予測モデル

文　　献

阿部重夫　2008　パターン認識のためのサポートベクトルマシン入門－Ⅱ：多クラスSVM．システム制御情報学会誌，**52**(9)：340-345

相見霊三・村上　高・藤巻和子　1956　水稲の登熟機構に関する生理的研究．日作紀，**25**：124-127

赤穂昭太郎　2008　カーネル多変量解析．岩波書店，東京，98頁

荒木雅登・山本富三・満田幸恵　2006　標準温度変換日数と生育診断による暖地水稲の窒素吸収量推定法．土肥誌，**77**(2)：191-194

Cataldo, D. A., L. E. Schrader and V. L. Youngs 1974　Analysis by digestion and colorimetric assay of total nitrogen in plant tissues high in nitrate. Crop Science **14**: 854-856

中鉢富夫・浅野岩夫・及川　勉　1986　葉緑素計による水稲（ササニシキ）の窒素栄養診断．土肥誌，**57**：190-193

森　静香・横山克至・藤井弘志　2010　山形県の庄内地域における登熟期の葉色診断による産米の玄米タンパク質含有率別仕分け法．日作紀，**79**：113-119

農業機械学会　1996　生物生産機械ハンドブック．コロナ社，東京．790-791頁

農林水産省　2011　一等米比率の推移及び平成22年産水稲うるち玄米の検査結果（平成23年1月31日現在）．In http://www.maff.go.jp/j/study/suito_sakugara/h2203/pdf/ref_data2-4.pdf　農林水産省'，農林水産省，東京

大山卓爾・伊藤道秋・小林京子・荒木　創・安吉佐和子・佐々木修・山崎拓也・曽根久美子・種村竜太・水野義孝・五十嵐太郎　1991　硫酸－過酸化水素分解法による植物，堆肥中に含まれるN，P，Kの分析．新大農研報，**43**：111-120

Platt, J.C.　1999　Fast Training of Support Vector Machines Using Sequential Minimal Optimization. *In" Advances in Kernel Methods: Support Vector Learning"*, ed. by B. Schölkopf, C. Burges and A. Smola, The MIT Press, Massachusetts, pp.185-208

Platt, J. C., N. Crixtianini and J. Shawe-Tailor　2000　Large Margin DAGs for Multiclass Classif Advances in Kernel Methods Support Vector Learning ication, *Advances in Neural Information Processing Systems*, **12**: 547-553

佐藤　庚・稲葉健吾・戸沢正隆　1973　高温による水稲の稔実障害に関する研究：第1報 幼穂形成期以降の生育時期別高温処理が稔実に及ぼす影響．日作紀，**42**(2)：207-213

佐藤　庚・稲葉健吾　1976　水稲の稔実障害に関する研究：第5報　稔実期の高温による籾の炭水化物受入れ能力の早期減退について．日作紀，**45**：156-161

田中明男・若松謙一・大内田　真　2010　暖地早期水

稲における日照不足が玄米品質に及ぼす影響．日作九支報，**76**: 9-11
田中 慶・木浦卓治・杉村昌彦・二宮正士・溝口 勝 2011 SIMRIWを利用した水田栽培可能性予測支援ツール．農業情報研究，**20**(1): 1-12
若松謙一・田中明男・上薗一郎・佐々木 修 2006 水稲の暖地早期栽培における登熟期間の遮光処理が収量，品質，食味に及ぼす影響．日作九支報，**72**: 19-21

Summary

We evaluated the support vector machine (SVM), a supervised learning method used for pattern recognition, as a fundamental analysis method for building prediction models of rice yield and quality. First, prediction models of patterns regarding the yield and protein content of brown rice were built using a directed acyclic graph SVM (DAGSVM), which is a multiclass classifier. These models predict patterns of yield and protein content on the basis of seven variables, including the state of rice plant after heading (nitrogen nutrition and yield capacity) and meteorological environment (air temperature and solar radiation) after the late spikelet initiation stage (i.e., 14 days before heading). Data used for building the models were obtained from surveys of 47 paddy fields conducted in 2009 in the village of Hoshino, in 2009 and 2010 at the experimental farm of Kyushu University, and in 2010 in the cities of Itoshima and Fukuoka. The Hinohikari cultivar was grown in all the surveyed fields. Next, the validity of the SVM as an analytical method for building prediction models was evaluated in terms of the classification rate and an adjustment function of generalization. The classification rates of yield and protein content were found to be relatively high, i.e., 85.1% and 76.6%, respectively. Further, it was confirmed that an adjustment function of generalization with soft margin was effective in training the data that were not linearly separated. The results indicated that SVM, with high accuracy and high generalization performance, is an effective method for building prediction models of rice yield and quality that are affected by various factors.

Key words: Prediction model, Protein content, Rice production, Support vector machine, Yield