２００１

遺伝的アルゴリズムを用いた波長領域選択手法の開発

○河村智史、荒川正幹、船津公人
東京大学大学院工学系研究科（〒113-8656 東京都文京区本郷 7-3-1）

【はじめに】近年、近赤外分光法を用いた研究が注目を集めている。近赤外とは 800~2500 nm 付近の波長領域を指し、古くから分子構造の推定などに用いられてきた赤外域に比べて、研究の遅れている波長領域であった。しかし、20 世紀後半に起こったケモメトリックスの発展によって、スペクトルデータに数学的手法や統計学的手法が適用されるようになり、近赤外分光法においても飛躍的な進歩が見られた。その結果、近年では近赤外光の非破壊性や非接触性などの利点が注目され、農業や医療をはじめとした幅広い分野の研究に応用されている。

しかしながら、近赤外分光法はそのデータ解析における方法論が体系的に確立されたわけではなく、ケモメトリックスの適用を通した更なる発展を必要としている。中でも近赤外スペクトルを用いたモデリングにおいては、スペクトルの持つ共線性が大きな障害となっており、オーバーフィッティングを避けた予測精度の高いモデルを構築する新しいデータ解析法の開発が求められている。

本研究ではこのような背景をふまえ、近赤外スペクトルを用いたモデリングにおける新しい領域選択手法の開発を行った。また、この手法を実際のデータへ適用し、その有用性の確認を行った。

【手法】

1. PLS 法

PLS (Partial Least Squares) 法[1]は MLR (Multiple Linear Regression)法を拡張した線形重回帰分析手法である。スペクトルデータのように、説明変数間に強い共線性がある場合や、データ数よりも説明変数の数が多い場合は、MLR 法でのモデル作成は困難となるが、そのような場合でも PLS 法を用いることで、信頼性の高い予測的なモデルを作成することが可能である。

2. GAPLS 法

GAPLS 法[2]は生物の進化を模倣した最適化手法である遺伝的アルゴリズム(GA:Genetic Algorithm)[3]を用いた変数選択法であり、PLS モデルの予測精度を表す Q^2 の値を最大化するような説明変数の組を選び出す。

3. GAWLS 法

本研究では、スペクトルデータにおいて効率的に変数選択を行うことのできる新しい変数選択法として、GAWLS(Genetic Algorithm-based WaveLength Selection)法を考案した。GAWLS 法は GA を用いてモデルの Q^2 を大きくする説明変数の組を領域単位で選び出す変数選択法である。Figure1 に GA 計算におけるそのアルゴリズムを示す。GAWLS 法では二つの実数値を用いて一つの波長領域を表現することで、モデル作成の際に必要となる重要な波長を領域単位で選択することができる。染色体の適合度は、染色体に表現された波長領域の全ての組み合わせについて PLS 解析を行い、得られた Q^2 の中で最大のものを採用している。これにより、必要な領域数がわからない場合においても GA 計算の効率を上げ、精度の高い近似解を求めることができる。スペクトルデータでは、波長間の相関が非常に高く、構造の特徴が幅をもったバンドで表されるため、GAWLS 法を用いて変数を選択することにより、オーバーフィッティングに陥ることなくモデルの改良が行え、さらには対象とする目的変数と関係を持つ波長領域の発見が期待できる。

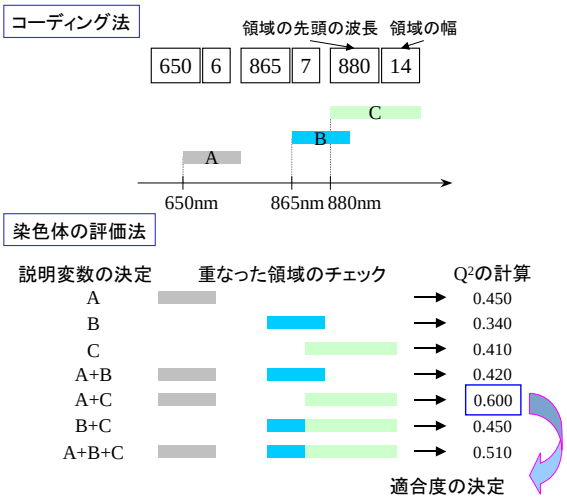

Figure1. GAWLS 法のアルゴリズム

【結果と考察】

解析データ

GAWLS 法の有用性を確認するため、精密農業における近赤外スペクトルデータに本手法を適用した。使用データは、土中光センサを備え付けたトラクターをある農地で走らせて測定した土壌の近赤外反射スペクトルと、土壌分析によって得られた土壌パラメータ(水分量、有機物量など)である。土壌の近赤外スペクトルから土壌パラメータを予測するモデルの構築を行った。

土壌水分量予測モデルの作成

説明変数を近赤外スペクトル、目的変数を土壌中の水分量として PLS 解析を行った。スペクトルのノイズ処理には、各種前処理法の比較検討により、高次の多項式で近似を行う Savitzky-Golay 法とセンタリングを採用した。その結果 R^2=0.724、Q^2=0.645、最適成分数は 6 となった。次に、モデルの予測精度の向上を目指し、従来手法である GAPLS 法と本研究で考案した GAWLS 法による変数選択を行った。GAPLS 法においては、世代数 1000、染色体数 50、GAWLS 法においては、世代数 500、染色体数 50 として計算を行い、充分な収束が得られていることを確認した。この計算を 20 回繰り返し、選択される波長の傾向を調べた。Figure2 と Figure3 にその結果を示す。

GAPLS 法を用いた場合、Q^2 は最大で 0.777 まで上昇し一見モデルの改良が行えたように見えるが、20 回の計算結果にばらつきが見られることからオーバーフィッティングに陥っている可能性が考えられる。これに対し GAWLS 法を用いた場合は、Q^2 は 0.732 まで上昇し、そのばらつきも小さかった。また選択される領域には一貫した傾向が見られ、H_2O の第一倍音である 1450 nm を含む領域が選択されていることから、水分量予測モデルとしての妥当性が確認できた。

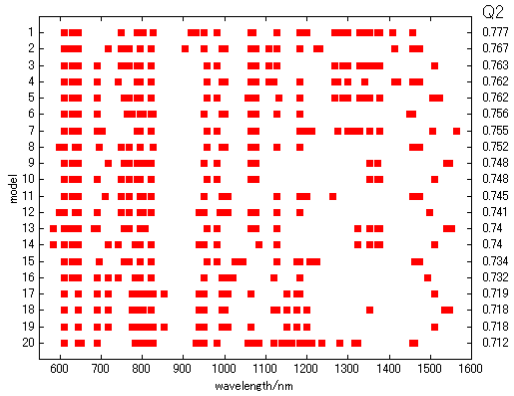

Figure2. GAPLS 選択波長分布

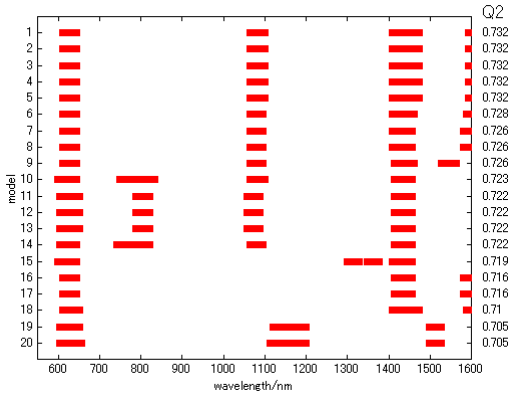

Figure3. GAWLS 選択波長分布

【まとめ】近赤外スペクトルデータのモデリングにおいて、目的変数と関係性の高い重要な波長を領域単位で選択する GAWLS 法を提案した。スペクトルデータは変数間の共線性が強く従来の変数選択手法ではモデルの改良を行うことは難しいが、GAWLS 法を用いることでスペクトルデータにおいても効率的に変数選択を行うことが可能になる。本研究ではこの手法を精密農業データへ適用し、その有用性の確認を行った。その結果、変数選択を行わなかった場合は、R^2=0.724 , Q^2=0.645 であったのに対し、GAWLS 法を用いることによって Q^2=0.732 まで改良され、従来使用されてきた変数選択手法である GAPLS 法を用いた場合よりも、安定的に高い予測精度を持つモデルを作成することができた。

【謝辞】本研究を行うにあたり、スペクトルデータを提供していただいた東京農工大学澁澤研究室に感謝いたします。

【参考文献】

[1] S. Wold, M. Sjostrom, L. Eriksson, Chemome. Intell. Lab. Syst., 58, 109-130 (2001).
[2] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci., 37, 306-310 (1997).
[3] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning., Addison Wesley, 1989.