4M-03

インク開発支援のための主成分分析と遺伝的プログラミングによる数理モデルの導出

及部 礼成[1]　安藤 大地[1]　恩田 陽人[2]　林 真一郎[2]　向井 智彦[1]
首都大学東京システムデザイン学部[1]　東洋インキ SC ホールディングス[2]

1. 序論

インクの調合や開発は職人の感覚や開発者の手計算によって行われていた．具体的には，複数の波長で構成した光を当て，吸収されずに透過した光の強さを測り，インクの材料選択を行う光透過の実験である．計測機器では，光の波長は 1000 帯域に分割して計測する．これらの帯域は，互いに相関関係が想定されている．人の手でこのような膨大で複雑な数値を計算するのは極めて困難である．そのため，新しいインクを作るには，開発者たちは毎回この実験を行わなければならないという問題があった．このような膨大な数値を分析し，数理モデル化することで，実験を毎回行う必要がなくなり，開発の効率を上げることができる．また，得られた数理モデルの規模が十分に人間が観察できるものであれば，開発する技術者のアイディアの創発につながる．本研究では，材料選択の際の光透過の実験とそのデータについて，主成分分析[1]を用いて次元数削減された潜在空間内において，遺伝的プログラミング（GP）[2]を適用することで，説明変数と目的変数の関係を記述する数理モデルを導出する（図 1 参照）．前述の技術者の観察を目的としているため，主成分分析にかける前の段階でデータを大きく次元圧縮する必要がある．

そこで本研究では，データの特徴を捉えたまま隣り合う説明変数の平均値を出す対話的次元圧縮で変数を減らす対話的次元圧縮を行うことで，変換行列が大きくなりすぎることを防ぐ．また，GP を用いた探索では，探索の際の関数の複雑化してしまうという問題点がある．そこで，GP と局所探索を組み合わせる[3]ことで，探索の際の関数の複雑化を防ぎ，全て変数が含まれるようになる．本手法を，実データを用いて検証した結果，高い近似精度が得られた．

図 1　提案手法の手順

2. 提案手法

2.1 対話的次元圧縮と主成分分析による次元圧縮

今回 750 以上の帯域の実験データは正確な観測が難しいことから除く．次元圧縮は図 1 中の(1)と(2)の二段階に分けて行う．第一段階の図 1 の(1)はデータの解像度を人間が対話的に下げていく．まずカラーメッシュ化するときに，データの特徴が失われないために，帯域ごとにカラーメッシュの最大値を調整して，人間が目視しやすいようにする．人間がデータの特徴を失われていないと感じるところまで，隣同士の帯域の値の平均をとる．このデータの場合，１２まで平均をとった状態が,「データの特徴が目視で失われない限界」だった．カラーメッシュから 450~700 の帯域がデータの特徴を顕著に示していることがわかる．そこで，450~700 の帯域に注目し，平均する近傍の数を決定する．

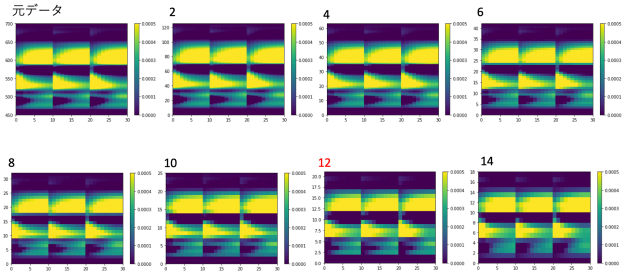

図 2　450~700 の帯域の元データと平均化したデータ

12 近傍を平均したデータに比べ，14 近傍を平均したデータは 450~500 の帯域に存在する谷の特徴がなくなっている．このことから，12 近傍の平均を用いることとした．（その後，後述する数理モデルの導出で，14 近傍以上を平均したデータの探索効率が悪くなっていることから，この対話的次元圧縮手法は，ある程度妥当性があると考えられる．）この対話的次元圧縮により，54 個まで変数を減らすことに成功した．次に，第二段階の図 1 中の(2)でさらに次元圧縮を行うことで，変数を 3 次元まで減らした．

2.2 数理モデル

モデル構築は GP を用いた関数同定の手法を用いて行う．モデルの近似精度の評価は決定変数とグラフを用いる．決定変数が高く，観測値とモデルのグラフと近似していれば

Derivation of a mathematical model using principal component analysis and genetic programming for ink development support
1 Tokyo Metropolitan University.
2 TOYO INK SC Holding

モデルの精度は高いと言える．

光の透過量を Y，第一主成分を X_0，第二主成分を X_1，第三主成分を X_2 とし，GP でモデルを構築したところ以下のようなモデルになった．

モデル：$173.771055355103*X_0 - 173.771055355103*X_2 + 0.728$・・・・・・・・・・・・・・・・・・・・・・(1)

テストデータの決定係数が高くなっている上，テストデータのグラフもほとんど近似している．(表 2 参照)

しかし，X_1 がモデル内に含まれていないため，局所探索により，さらに近似精度が高いモデルを探索する．ランダムサーチは解の全ての可能性の空間を考え，その中から無作為に選び出した解が条件に合致するかを判定する方法である．この方法を用い，X_1 を含んだモデルを探索する．GP で生成したモデル(1)に以下のような式をランダムサーチで生成し項として加える．

$+aX + \exp(bX + c)$・・・・・・・・・・・・・・・・・(2)

a，b，c は 0 から 1 までのランダム実数であり，+の位置には，+か-が入り，X は $1, X_0, X_1, X_2$ のどれかが入る．なお，生成した式の中に X_1 が含まれなければ，除外するものとする．このようにしてランダムサーチした結果，式(3)の通りになった．

(2) に当てはめた式 $:+0.6059634794841777*1-\exp(0.548620204595044*X_1-0.5031080123343032)$・・・・(3)

ランダムサーチを行なった結果，GP の時よりも決定係数が高くなっている．(表 2 参照)

次に山登り法を用いて，さらに近似精度が高くなる定数を探索する．今回定数が 3 つあるため，定数 b→a→c の順に 0.00001 ずつ決定係数が高くなるように増やすかもしくは減らした結果，式(4)に示す通りになった．

(2) に当てはめた式：$+0.6059334794841779*1-\exp(20.170920204595244*X1-0.5031280123343032)$・・・・(4)

表 1 局所探索で作成したモデルの決定係数

GP		GP＋ランダムサーチ		GP+ランダムサーチ＋山登り法	
訓練	テスト	訓練	テスト	訓練	テスト
0.998462	0.9939743	0.99846293	0.99397 4336	0.99861 7472	0.99406 9585

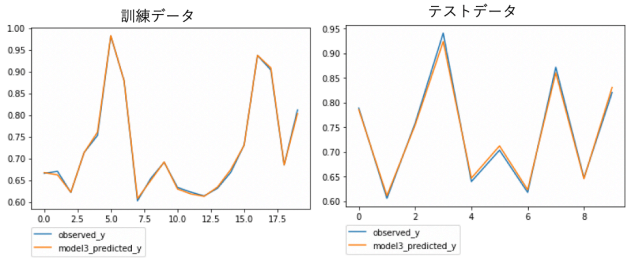

図 3　GP+山登り法で作成したモデルの予測と観測データ

GP＋ランダムサーチの時よりも局所探索を行うことで，決定係数は高くなっている．（表 2 参照）このことから，GP と局所探索を組み合わせることにより，近似精度をさらに高めることができた．

2.3　訓練データ以外に適用した時の性能

訓練データで導出したモデルと主成分への変換行列を他のデータにも適応した際の性能を検証する．12 近傍を平均して次元圧縮した他データを，訓練データで導出した主成分への変換行列を用いて主成分にし，式(4)のモデルを適用した．その結果，以下の通りになった．なお今回用意した実データは，試行分のデータ（データ B）と 52 試行分のデータ(C データ)である．

表 2　実データに対する効果

	訓練	テスト	B	C
決定係数	0.998671	0.993821	0.995979	0.965420

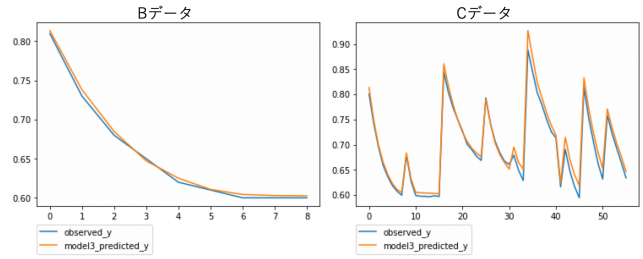

図 4　モデルの予測と観測データ

両方のデータとも決定係数が高く，かつグラフも観測値と近似しているため，近似精度は高いことがわかる．このことから，今回の手法で得た主成分への変換行列とモデルは汎用性が高いと言える．

3.　結論

本研究ではブラックボックス化したデータを技術者が観察し，考察できることを目的とした数理モデル導出方法を提案した．この手法は，他のデータにも訓練データで導出した主成分への変換行列とモデルを適用した結果，他のデータに対しても高い性能を持つことがわかった．しかし，今回，決定係数は高くなったものの，グラフを見ると一部近似できていないところも散見された．したがって，より数理モデル構築の精度を高めていくことが今後の課題である．

参考文献

[1] 塚本邦尊,山田典一,大澤文孝: 東京大学のデータサイエンティスト育成講座　Python で手を動かして学ぶデータ分析,株式会社マイナビ出版(2019)

[2] 伊庭斉志,佐藤泰介:システム同定アプローチに基づく遺伝的プログラミング

[3]関岡哲也,船曳信生,東野輝夫:遺伝的プログラミングを用いた関数合成アルゴリズムの改良法提案,電子情報通信学会論文誌　D-I,vol J83-D-INo.4,pp.407-417,200.