Vol.2022-MPS-137 No.20
2022/3/4

確率的スキーマ貪欲法を用いた自動機械学習

牧野 寛也[1,a)]　北 栄輔[1,b)]

概要：確率的スキーマ貪欲法 (Stochastic Schemata Exploiter, SSE) は進化的計算手法の一つである．多くの進化的計算手法と同様 SSE の遺伝子はバイナリ値で表現されるが，ハイパーパラメータ最適化に応用しようとすると離散値やラベル値も表現しうるべきである．そこで，初期集団生成，スキーマ抽出，個体生成，突然変異，および世代交代についてアルゴリズムを拡張する．拡張したアルゴリズムを用いて，アンサンブル手法の一種であるスタッキングモデルを最適化する．構造を制御する変数であるモデルシンボルを導入し，シンボルを他のハイパーパラメータと同様に扱うことで，スタッキングの構造とハイパーパラメータを同時に最適化する．数値実験により，SSE は既存手法と比較して高い性能を持つことを確認する．また，プロセスの可視化を行い，本来ブラックボックスである最適化の過程を理解しやすくする．

1. はじめに

近年，機械学習はますます重要になりつつある．多くの機械学習手法はハイパーパラメータを持つが，知識や経験なしにそれを設定するのは容易ではない．本研究では確率的スキーマ貪欲法 (Stochastic Schemata Exploiter, SSE) を用いたハイパーパラメータ最適化手法を提案する [1]．SSE は Aizawa [2] によって 1994 年に提案された，進化的計算手法の一つである．Aizawa は，スキーマ貪欲法を「集団中に存在するスキーマの中で，サンプル平均のよいスキーマをさらにサンプルすること」と定義した．そのようにして生成されたスキーマから確率的に新たな個体を作成し，その個体集団を母集団としてを再びスキーマサンプリング，個体の生成を繰り返すことによって適応度の高い個体を得ようとするのが SSE である．SSE の長所は，アルゴリズムそのもののハイパーパラメータが少ない点，はやい収束特性を持つ点にある．そこで，本研究では SSE を HPO について適用することについて述べる[*1]．

2. 確率的スキーマ貪欲法

確率的スキーマ貪欲法のアルゴリズムは下記の通りである．ここで，M を個体数とし，$\{0, 1, *\}$ からなる文字列をスキーマとする．$*$ はその位置に $0, 1$ のどちらが来ても良いことを意味する．

(1) 初期世代として M 個の個体をランダムに生成する
(2) 終了条件を満たすまで以下を繰り返す
  (a) 各個体の適応度を計算する
  (b) 個体を適応度の降順に並べ，最上位の個体から半順序関係に従って個体部分集合を生成する．個体部分集合のうち，平均評価値の上位 M 個をリストに格納する
  (c) M 個の各個体部分集合において，共通スキーマを抽出する
  (d) 共通スキーマ上の $*$ を 0 か 1 でランダムに置き換えることで子個体を生成する
  (e) 突然変異操作を適用する
  (f) 生成された M 個の個体を次世代の個体とする

個体部分集合を全て得ようとすると，$2^M - 1$ 個にのぼる．2b で言及した半順序関係を利用した個体部分集合の生成は，平均評価値の上位 M 個の個体部分集合を効率的に取り出すための工夫である．その他詳細については Aizawa [2] を参照されたい．

3. 提案手法

3.1 SSE を用いたハイパーパラメータ最適化

SSE は他の多くの進化的計算手法と同様，バイナリ値によって個体を表現する．しかしながら，HPO に応用しようとすると離散値やラベル値も表現しうるべきである．

提案手法では，はじめに各遺伝子の候補を集合 $\Gamma_j\,(j = 1, 2, \cdots, L)$ で定義する．ただし，L は遺伝子長である．初期集団の各個体の遺伝子は，Γ に基づきランダムに選択する．合わせて，スキーマの定義も拡張する．共通スキーマ

[1] 名古屋大学大学院 情報学研究科
Graduate School of Informatics, Nagoya University
[a)] makino.hiroya.a3@s.mail.nagoya-u.ac.jp
[b)] kita@is.nagoya-u.ac.jp
[*1] https://github.com/macky168/sseopt にて提案手法のソースコードを参照可能

に作用素 $\mathcal{R}$ を適用することで新しい個体を生成する．$\mathcal{R}$ は集合を入力とし，全要素からランダムに選択した一つを出力する．

また突然変異は，要素 x'_j を対応する Γ_j からランダムに選択した要素 $\mathcal{R}(\Gamma_j)$ によって置換することで表現する．さらに突然変異率に関して，ランクベース突然変異を提案する．突然変異率が高い場合良い個体を破壊する可能性があり，低い場合集団の多様性を十分に保つことができない．これらの課題を解決するのがランクベース突然変異である．

- **通常の突然変異**：突然変異率 P_{m} は固定される
- **ランクベース突然変異**：突然変異率は共通スキーマの順位，すなわち元となる個体部分集合の順位による．順位 i $(i = 1, 2, \cdots, M)$ のスキーマから作成された個体は，確率 $p(i)$ で突然変異する．P_{mmax} を所与の定数として，

$$p(i) = \frac{i-1}{M} \cdot P_{\mathrm{mmax}}. \tag{1}$$

3.2 遺伝子表現

提案手法では，スタッキングの構造とハイパーパラメータを同時に最適化する．ここで，モデルシンボルを定義する．シンボルは，各弱学習器のパラメータの先頭に置かれる．シンボルの値は 0 か 1 を取り，0 の場合弱学習器の出力はスタッキングで使用されない．このシンボルを他のパラメータと同様に扱うことで，スタッキングの構造とパラメータを同時に最適化する．**図 1** は遺伝子表現を示す．

4. 数値実験

数値実験では，UCI リポジトリから取得した Abalone データセット，及び Gas Turbine CO Emission Dataset データセットを使用する．Neural Network，XGBoost，Random Forest，K-nearest Neighbors Regression を弱学習器とするスタッキングモデルを最適化する．

Abalone データセットにおける結果を **図 2** に，Gas Turbine CO Emission データセットにおける結果を **図 3** に示す．横軸が試行回数，縦軸が 10 回の実験における $\mathcal{R}^2$ 値の平均である．

いずれの結果においても，SSE が最も良い値に収束している．また，ランクベース突然変異は通常の突然変異よりも良い値を示した．

また，**図 4** は最適化の過程を図示した一例である．この例においては，XGBoost と Random Forest が初期に 1 へ収束したこと，K-nearest Neighbors Regression は一度 1 に落ち着いたものの突然変異が効果的に働いて最終的に 0 に収束したことなどを読み取ることができる．

5. おわりに

本研究では SSE を利用したハイパーパラメータ最適化手法を提案し，スタッキングモデルのハイパーパラメータ最適化に適用した．数値実験において，提案手法を GA，TPE，CMA-ES，Random Search と比較したところ，既存の手法よりも優れた結果を示した．このことにより，提案手法の有効性を確認した．また提案手法は，ハイパーパラメータ最適化過程の可視化に有効であることを示した．今後は，実社会データに対する機械学習の適用を通して，有効性をさらに確認していきたい．

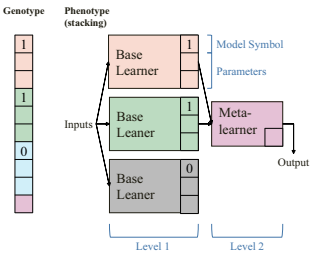

図 1 遺伝子表現．

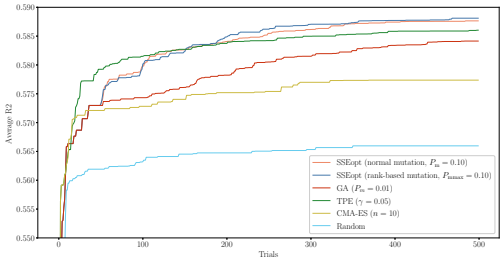

図 2 Abalone データセットにおける結果．

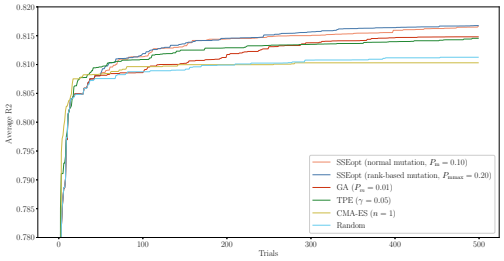

図 3 Gas Turbine CO Emission データセットにおける結果．

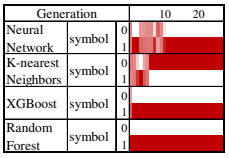

図 4 SSE の最適化過程の可視化．

参考文献

[1] Makino, H. and Kita, E.: Stochastic schemata exploiter-based AutoML, *Proceedings of the 2021 international conference on data mining workshops (ICDMW)*, pp. 238–245 (2021).

[2] Aizawa, A. N.: Evolving SSE: A stochastic schemata exploiter, *Proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence*, pp. 525–529 (1994).