重力探索アルゴリズムと遺伝的アルゴリズムを組み合わせた特徴量選択手法の提案

乾遼太郎[†1]，大場春佳[1]， 生田目崇[1]

キーワード：重力探索アルゴリズム，遺伝的アルゴリズム，特徴量選択

1. 研究背景と目的

近年，機械学習の実務・研究において扱われるデータは大規模化・高次元化しており，冗長あるいは無関係な特徴量を多数含むことが多い．このような高次元データでは，計算量の増大や過学習，モデル解釈性の低下といった問題が生じるため,特徴量選択が重要となる[1]．特徴量選択は組合せ最適化問題として定式化され，厳密解探索が困難なNP 困難問題であることから[2]，近似解を効率的に探索可能なメタヒューリスティクスが広く用いられてきた．

しかし，単一のメタヒューリスティクスでは局所解への停滞や探索と活用のバランスに課題が残るため，異なる探索特性を持つアルゴリズムを組み合わせたハイブリッド化が注目されている．既存研究では，重力探索アルゴリズム（GSA）[3]に遺伝的操作を導入した HGSA [4]が提案されているが，交叉による追加評価が計算コスト増大を招く可能性や，多様性維持の観点で改良の余地が指摘されている．

本研究では，GSA と遺伝的アルゴリズムの一種であるCHC [5]を統合したハイブリッド特徴量選択手法 GCHC を提案する．GSA の収束能力を活かしつつ，HUX 交叉や近親交配回避，リスタート機構を有する CHC を組み込むことで，局所解停滞の回避と探索多様性の維持を図り，高次元特徴量選択における探索効率と解品質の両立を目指す．

2. 既存手法と提案手法

2.1 重力探索アルゴリズム(GSA)

重力探索アルゴリズムは，ニュートンの重力法則と質量相互作用に着想を得たメタヒューリスティクス最適化アルゴリズムである．各エージェント（解）を物体と見なし，解の性能を質量として扱う．質量の大きい（良い）解が他の解を引き寄せることで，集団全体がより良い解の方向へ移動しながら探索を行う．

2.2 ハイブリッド重力探索アルゴリズム(HGSA)

GSA 単体は収束性に強みを持つ一方，多様性の低下により局所解へ陥る可能性がある．このため，交叉と突然変異の遺伝的操作を取り入れ，多様性を補うハイブリッド GSA が提案されている．

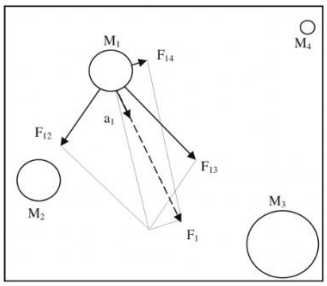

図 1　Gravitational Search Algorithm [3]

HGSA では，GSA に基づく探索過程に遺伝的アルゴリズムの交叉および突然変異操作を組み込むことで，探索多様性の維持を図る．交叉操作では，集団内のグローバル最良エージェントと適応度の低いエージェント間で特徴量選択ベクトルの一部を交換し，有効な特徴量構造を保持しつつ新たな組合せを生成する．また，探索が停滞した場合には突然変異を適用し，選択状態を反転させることで探索領域を拡張し，局所解からの脱出を促進する．

2.3 提案手法 (GCHC)

提案手法 GCHC は，GSA と CHC アルゴリズムを統合したハイブリッド特徴量選択手法である．GSA による物理法則ベースの収束機構と，CHC による多様性維持および強力な探索能力を組み合わせることで，探索と収束のバランス改善を狙う．GCHC のフローは以下の 4 ステップからなる．

1. **CHC ステップ（HUX 交叉・交叉世代エリート選択）**
2. **GSA ステップ（質量・重力に基づく更新）**
3. **総合選抜**
4. **リスタート処理**

まず CHC ステップにおいて，ハミング距離制約の下で HUX 交叉を適用し，親個体と子個体を統合したプールからエリート選択を行い，多様性を維持しつつ良解を生成する．次に GSA ステップでは，既存手法と同様に計算を行う．その後，CHC および GSA によって得られた個体群を統合し，適応度に基づく総合選抜により次世代の個体群を構成する．さらに，一定の反復で最良解が更新されない場合には停滞と判断し，部分的リスタートを行い，多様性を再導入し局所解からの脱出を図る．

†1　中央大学

3. 検証方法

本研究では，提案手法を検証するために特徴量全選択モデル，粒子群最適化(PSO)，重力探索アルゴリズム(GSA)，ハイブリッド重力探索(HGSA)を比較手法として用いる．
使用するデータはサンプル数 60~200，特徴量数 2,000~10000 の高次元二値分類データセット（Leukemia, Colon Cancer, DLBCL, Arcene）を用いて，提案手法の有効性を検証した．
予測性能とより少ない特徴量数の両立を目指すにあたり，アルゴリズムの目的関数は

$$0.99 \times ErrorRate + 0.01 \times \frac{\#Selected\ Features}{\#All\ Features}$$

とする．

特徴量選択後の予測モデルにはロジスティック回帰を用いる．正則化項にはL_2正則化を用い，正則化係数Cをハイパーパラメータとして調整する．分類性能の評価指標は Accuracy, Precision, Recall, F1-score を用いる．また特徴量選択アルゴリズムと予測モデルのハイパーパラメータ調整を含めた性能評価のため，Nested Cross-Validation を採用する．外側ループではデータを$K = 5$分割し，汎化性能の推定を行う．内側ループでは3分割し，特徴量選択アルゴリズムとロジスティック回帰の正則化係数Cを同時に最適化する．

4. 検証結果

以下が各データセットにおける結果である．

表 1 Leukemia

MODEL	ALL	PSO	GSA	HGSA	GCHC
Features	7129	3238.2	3348.4	3232.6	**3052.6**
Accuracy	0.9314	0.9448	0.9448	0.9448	**0.9590**
Precision	0.9358	0.9448	0.9448	0.9448	**0.9614**
Recall	0.9389	0.9489	0.9489	0.9489	**0.9600**
F1	0.9279	0.9409	0.9409	0.9409	**0.9558**
Time[s]	**27.54**	4037.19	3755.18	30825.30	13865.40

Leukemia では，提案手法 GCHC が各指標で最良となった．また，PSO/GSA/HGSA と同程度以上の精度向上を保ちつつ，選択された特徴量数はそれらより少なく，性能と削減のバランスが改善した．GCHC の計算時間は HGSA よりも短いが，その他の手法と比べると長い．

表 2 Colon Cancer

MODEL	ALL	PSO	GSA	HGSA	GCHC
Features	2000.00	872.60	882.20	787.80	**695.20**
Accuracy	0.7910	0.7897	0.7910	0.7897	**0.8385**
Precision	0.7984	0.7920	0.7984	0.7819	**0.8340**
Recall	0.8125	0.8125	0.8125	0.8050	**0.8500**
F1	0.7827	0.7819	0.7827	0.7814	**0.8295**
Time[s]	**7.02**	1173.02	1436.68	7925.02	4331.90

Colon Cancer においても GCHC が最良の精度を示し，さらに選択された特徴量数も他のメタヒューリスティクスより少ない結果となった．

表 3 DLBCL

MODEL	ALL	PSO	GSA	HGSA	GCHC
Features	5469.00	2517.20	2598.20	2477.80	**2318.00**
Accuracy	0.8717	0.8842	**0.8850**	0.8708	0.8583
Precision	0.8744	0.8800	**0.8816**	0.8600	0.8494
Recall	0.9136	0.9220	**0.9227**	0.9129	0.9053
F1	0.8642	0.8757	**0.8767**	0.8598	0.8456
Time[s]	**20.32**	3763.50	3722.76	25749.90	12490.80

DLBCL では PSO/GSA が ALL を上回ったが，GCHC は精度指標で相対的に劣る結果となった．一方，GCHC は選択された特徴量数が最少となった．以上 2 つのデータセットの計算時間については Leukemia と手法間の相対的な大小関係は同様であった．

表 4 Arcene

MODEL	ALL	PSO	GSA	HGSA	GCHC
Features	10000.00	4769.40	4863.00	4752.20	**4542.00**
Accuracy	0.7800	0.7800	0.7700	0.7700	**0.8100**
Precision	0.7958	0.7936	0.7830	0.7944	**0.8298**
Recall	0.7878	0.7822	0.7759	0.7774	**0.8143**
F1	0.7771	0.7762	0.7673	0.7650	**0.8074**
Time[s]	**63.45**	14069	19714	107691	58668

Arcene では，GCHC が各指標で最良となった．また，選択された特徴量数も他の手法と比べ少ない．計算時間については Leukemia と手法間の相対的な大小関係は同様であったが，他の 3 つのデータと比べると全体的に長くなった．

5. 考察

実験結果より，提案手法 GCHC は 4 データセットのうち 3 データセットにおいて，分類性能と特徴量削減の両者において既存手法を上回る有効性を示した．これらのデータセットでは，GSA による収束特性と CHC による多様性維持機構が補完的に機能したと考えられる．一方，DLBCL では他手法と比較して性能が低下する結果となった．これは，特徴量削減による情報損失が性能に影響した可能性がある．また，GCHC は PSO や GSA と比較して計算時間が増加する傾向があるものの，HGSA よりは大幅に短縮されており，性能と計算コストのバランスは一定程度改善されている．

参考文献

[1] Guyon, I. and Elisseeff, A., “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, Vol.3, pp.1157-1182, 2003.
[2] Xue, B., Zhang, M. and Browne, W. N., “A survey on Evolutionary Computation Approaches to Feature Selection”, *IEEE Transactions on Evolutionary Computation*, Vol.20, No.4, pp.606-626, 2016.
[3] Rashedi, E., Nezamabadi-pour, H. and Saryazdi, S., “GSA: A Gravitational Search Algorithm”, *Information Sciences*, Vol.179, No. 13, pp. 2232-2248, 2009.
[4] Taradeh, M., Mafarja, M., Heidari, A. A., Faris, H., Aljarah, I., Mirjalili, S. and Fujita, H., “An evolutionary gravitational search-based feature selection”, *Information Sciences,* Vol.497, pp.219-239, 2019.
[5] Eshelman, L.J., “The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination”, *Foundations of Genetic Algorithms*, pp.265-283, 1991.