並列分散型多目的ファジィ遺伝的機械学習を用いた

アンサンブル識別器設計

第 8 グループ 面﨑 祐一

# はじめに

ファジィ識別器は言語的に解釈可能なルール集合で構成されるため，どのようにデータを識別しているのかが理解可能であるという特徴をもつ．しかし，識別性能の高さと解釈性能の高さにはトレードオフの関係があるため，どちらも同時に最適となる識別器の獲得は困難である．そこで，ファジィ識別器の設計に進化型多目的最適化手法を用いた多目的ファジィ遺伝的機械学習 (Multiobjective Fuzzy Genetics-Based Machine Learning: MoFGBML) [1] が提案されている．近年では，大規模なデータに対する機械学習の発展が期待されている．一方で，MoFGBMLは高い識別性能を持つが，大規模なデータに適用する際に膨大な計算時間を必要とする問題がある．先行研究 [1] では，Island型の並列分散手法をMoFGBMLに適用した計算時間の短縮が提案されている．

[1] の並列分散型MoFGBMLでは部分学習用データへの過学習を防ぐために，部分学習用データの交換操作と部分個体群の移住操作が行われているが，その結果，識別性能を重視した識別器の獲得が困難であるという課題がある．そこで，識別性能の高い識別器を獲得するため，各部分個体群から抽出した弱識別器で構成されるアンサンブル識別器を設計する．本研究では，[1] で行われていた交換操作，移住操作は適用せずに学習を行う．これにより，各島で独立な進化型多目的最適化が行われるため，抽出した弱識別器の間には高い多様性が期待できる．これらの弱識別器でアンサンブル識別器を構成することで，識別性能の高い識別器の獲得を図る．

# 多目的ファジィ遺伝的機械学習

## ファジィ識別器

*n*次元*M*クラスのパターンが*m*個与えられたパターン識別問題に対して，ファジィ集合を前件部とする以下のIf-thenルールを用いてファジィ識別器を設計する．ある未知パターンは**x** = (*x*1, …, *xn*) のように表され，*xi*は第*i*次元 (*i* = 1, 2, …, *n*) における属性値を表す．

Rule *R*: If *x*1 is *A*1 and ... and *xn* is *An*

then Class *C* with *CF*  (1)

**A** = (*A*1, …, *An*)は条件部ファジィ集合，*C*は結論部クラス，*CF*はルールの重みを表す．本研究では，条件部ファジィ集合として，2, 3, 4, 5分割の三角型のファジィ集合14種と，メンバシップ値として必ず1を返す ”don’t care” の合計15種類のファジィ集合を同時に用いる．学習用データを用いてルール重みと結論部クラスを決定する．未知パターンの推論は，適合度とルール重みの積が最大となるルールを勝者とする単一勝利ルールによって行う．

## 多目的ファジィ遺伝的機械学習

本研究では，代表的な進化型多目的最適化アルゴリズム (Evolutionary Multi-objective Optimization Algorithm: EMOA) であるNSGA-II [2] をFGBMLに適用する．これによって，識別器の誤識別率最小化とルール数最小化の2つの目的を同時に最適にするような識別器の探索を行う．以下にMoGBMLの手順を示す．

Step 1: 学習用データから初期個体群を生成し，初期個体群の評価を行う．

Step 2: 現個体群から遺伝的操作（交叉，突然変異操作）によって子個体群を生成する．

Step 3: 子個体を評価し，現個体群の世代更新を行う．

Step 4: 終了条件を満たさない場合，Step 2 へ戻る．

Step 5: 得られた個体群から，誤識別率最小化，ルール数最小化の2目的において互いに非劣な個体を選択する．

# 並列分散型MoFGBMLによるアンサンブル識別器

## 並列分散実装

先行研究 [1] において，MoGBMLの計算時間短縮のための方法としてIsland型の並列分散実装が提案されている．並列分散実装では，個体群と学習用データを分割し，部分個体群と部分学習用データのペアをそれぞれ一つのCPUコアに割り当てて進化型多目的最適化を行い，計算時間の短縮を図る．Island型並列分散実装では，一定間隔で最良個体の異なる島への移住操作と部分学習用データ集合の交換操作が行われる．これらの二つの操作により，部分個体群の部分学習用データへの過学習を防ぎ，汎化性能を向上させることが期待される．各島において学習終了後，全ての個体は一つの個体群に統合される．最後に，全ての個体は誤識別率とルール数の同時最小化における非優劣ランキングに基づいて評価される．

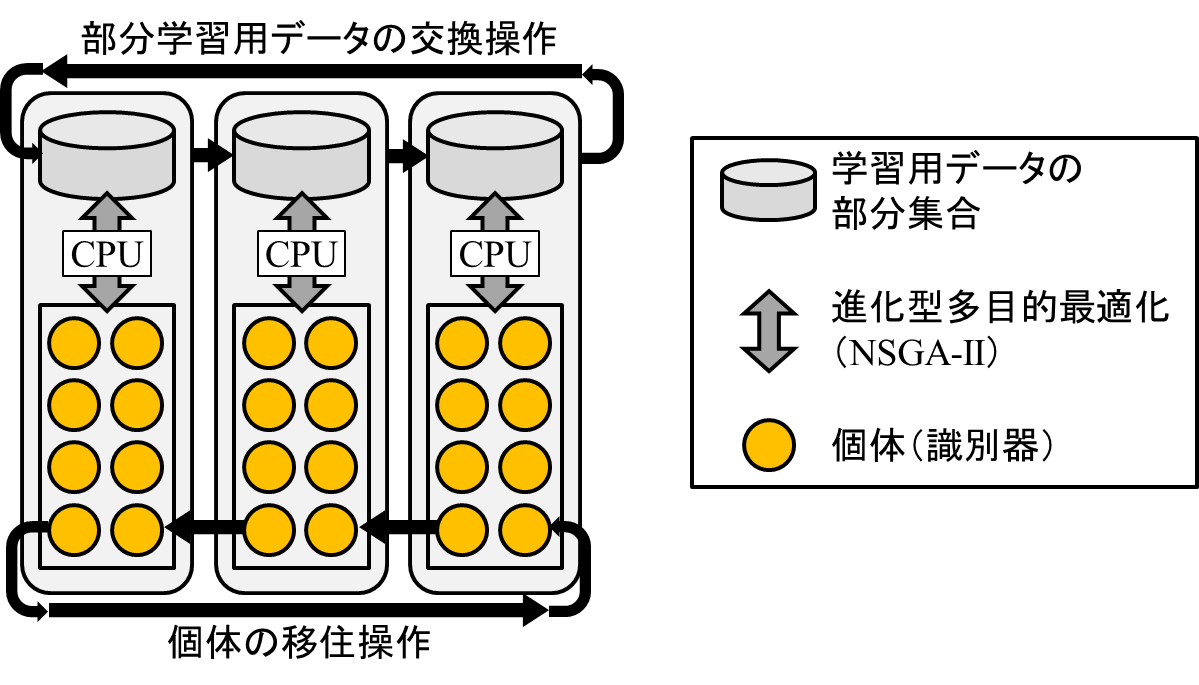


図1: 並列分散実装のモデル図 [1]

## アンサンブル識別器の設計

識別器の設計において，複数の弱識別器で構成されるアンサンブル識別器の設計が提案されている [3]．アンサンブル識別器は複数の弱識別器の識別結果で多数決を行う．得られた識別器集合の内，特定の未知パターンに対して識別性能が低くなる弱識別器が存在しても，他の弱識別器の識別結果を利用できる．そのため，単一な識別器と比較して，識別性能の向上が期待される．

本研究では，弱識別器の多様性向上のため，交換操作，移住操作を行わない並列分散型MoFGBMLによるアンサンブル識別器の設計を行う．この並列分散型MoFGBMLで得られた部分個体群から，島ごとに最良の識別器を抽出し弱識別器とする．各島は独立な進化型多目的最適化が行われるため，アンサンブル識別器を構成する弱識別器に高い多様性が期待できる．

また，アンサンブル識別器の識別性能の向上を目的とした重み付け多数決によるアンサンブル識別器を設計する．これは，各弱識別器の識別性能 (識別率) を重みとして与えた重み付け多数決による識別を行う．単純多数決と比較して，多数決結果が同票となる可能性を低くできる．

# 数値実験

## 数値実験設定

本研究で設計したアンサンブル識別器と [1] の並列分散実装で得られた識別器の識別性能を比較した．このとき，[1] で行う交換間隔と移住間隔はともに50世代とした．また，並列分割数の違いによるアンサンブル識別器の識別性能の変化についての実験を行った．数値実験にはKEEL-data set repositoryにより提供されている以下の2種類の実世界データを用いた（表1）．それぞれのデータの連続値を持つ属性においては，各属性値の最小値と最大値を用いて [0, 1] の範囲で正規化する．以下に数値実験における各種設定を示す．

|  |  |  |
| --- | --- | --- |
| 試行回数 | : | 30 (10-fold cross-validation x 3) |
| 終了条件 | : | 50,000 |
| 個体群サイズ | : | 300 |
| EMOA | : | NSGA-II |
| 並列分割数 | : | 3, 5, 7, 9 |
| 移住操作間隔 | : | 50世代間隔，なし |
| 交換操作間隔 | : | 50世代間隔，なし |

表1: 使用するデータセット

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Patterns | Attributes | Classes |
| Phoneme | 5404 | 5 | 2 |
| Satimage | 6435 | 36 | 6 |

## 数値実験結果

### 単一弱識別器によるアンサンブル識別器

本節では，各島の部分個体群から，全学習用データに対する誤識別率が最小の個体を弱識別器として抽出したアンサンブル識別器について実験を行った．重み付け多数決では，割り当てられた部分学習用データの識別率を重みとするサブデータ重み多数決と，全学習用データの識別率を重みとする全データ重み多数決の2種類の方法で実験した．加えて，学習用データに対して識別率が最も高い識別器を単一に選択した識別器の識別性能との比較を行った．

評価用データに対する各識別器の誤識別率の30試行した平均を表2，表3に示す．[1] と比較して提案手法の誤識別率が低い結果を太字で表している．また，最も低い誤識別率の結果に下線を引いている．

表2: 評価用データの誤識別率 (Phoneme, 単一弱識別器)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 島数 | 並列分散手法 [1] | 単一  識別器 | 単純 多数決 | サブデータ  重み多数決 | 全データ 重み多数決 |
| 9 | 17.28 | 18.14 | **17.13** | **17.14** | **17.14** |
| 7 | 17.06 | 17.89 | **16.75** | **16.76** | **16.75** |
| 5 | 16.19 | 17.09 | 16.27 | 16.27 | 16.27 |
| 3 | 15.44 | 16.38 | 15.52 | 15.53 | 15.53 |

表3: 評価用データの誤識別率 (Satimage, 単一弱識別器)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 島数 | 並列分散手法 [1] | 単一  識別器 | 単純 多数決 | サブデータ  重み多数決 | 全データ 重み多数決 |
| 9 | 14.58 | 16.48 | **14.25** | **14.21** | **14.24** |
| 7 | 14.31 | 16.05 | **14.26** | **14.23** | **14.27** |
| 5 | 13.96 | 15.03 | 14.05 | 14.00 | 13.98 |
| 3 | 13.57 | 14.34 | **13.26** | **13.24** | **13.16** |

表1, 2より，島数9, 7の場合，提案手法による誤識別率が最も低くなっていることが分かる．また，Phonemeに対しては単純多数決が最も低く，Satimageに対してはサブデータ重み多数決が最も低くなることが分かった．

### 非劣解集合弱識別器によるアンサンブル識別器

本節では，全学習用データの誤識別率最小化とルール数最小化の2目的に対して，NSGA-II [2] における非優越ランキングに基づいた非劣な識別器集合を弱識別器とする．このとき，抽出した非劣解集合を全て用いて多数決を行う．評価用データに対する各識別器の誤識別率の30回試行した平均を表4, 5に示す．

表4: 評価用データの誤識別率 (Phoneme, 非劣弱識別器)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 島数 | 並列分散手法 [1] | 単一  識別器 | 単純 多数決 | サブデータ  重み多数決 | 全データ 重み多数決 |
| 9 | 17.28 | 18.14 | 18.31 | 18.27 | 18.29 |
| 7 | 17.06 | 17.89 | 17.64 | 17.50 | 17.51 |
| 5 | 16.19 | 17.09 | 17.10 | 17.11 | 17.14 |
| 3 | 15.44 | 16.38 | 16.17 | 16.17 | 16.17 |

表5: 評価用データの誤識別率 (Satimage, 非劣弱識別器)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 島数 | 並列分散手法 [1] | 単一  識別器 | 単純 多数決 | サブデータ  重み多数決 | 全データ 重み多数決 |
| 9 | 14.58 | 16.48 | 17.00 | 16.37 | 16.49 |
| 7 | 14.31 | 16.05 | 16.71 | 16.06 | 16.12 |
| 5 | 13.96 | 15.03 | 15.88 | 15.39 | 15.41 |
| 3 | 13.57 | 14.34 | 13.95 | 13.65 | 13.70 |

表4, 5より，非劣解集合で構成された弱識別器では並列分散手法 [1] で得られた識別器と比べて識別性能が劣ることが分かった．

# おわりに

# 参考文献

1. Y. Nojima, Y. Takahashi, and H. Ishibuchi, “Application of parallel distributed implementation to multiobjective fuzzy genetics-based machine learning,” *Lecture Notes in Computer Science 9011: Intelligent Information and Database Systems – ACIIDS 2015*, Part I, pp. 462-471, Springer, Berlin, March 2015.
2. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Trans. on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, 2002.
3. L. Rokach, “Ensemble-based classifiers,” *Artificial Intelligence Review*, vol. 33, pp. 1-39, 2010.