

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 面崎 祐一

1. はじめに

ファジィ識別器は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言語的に解釈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化す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、ビッグデータなどの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。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]。

従来の並列分散型 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、獲得される識別器集合の多様性が劣る課題がある。本研究では、交換操作、移住操作を適用せずに多様性を考慮した識別器の探索を行う。また、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。各島では独立な進化型多目的最適化が行われるため、各島から抽出した弱識別器の間には高い多様性が期待できる。これらの弱識別器の多数決によってパターンを識別することで、識別性能の向上を図る。

2. 多目的ファジィ遺伝的機械学習

2.1. 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条件とする以下の 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 $\mathbf{x} = (x_1, \dots, x_n)$ のように表され、 x_i は第 i 次元 ($i = 1, 2, \dots, n$) における属性値を表す。

Rule R: If x_1 is A_1 and ... and x_n is A_n
then Class C with CF , (1)

R はファジィルール、 $\mathbf{A} = (A_1, \dots, A_n)$ 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し、未知パターンに対する適合度とルール重みの積が最大となる単一勝利ルールによって推論を行う。

2.2. 多目的ファジィ遺伝的機械学習

本研究では、個体の非優越ランキングに基づく進化型多目的最適化アルゴリズムである NSGA-II [2] を FGBML に適用し、識別器の誤識別率とルール数の 2 つの目的を同時に最小化するような識別器の探索を行う。以下に多目的ファジィ GBML 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: 各ルール集合に対して遺伝的操作 (交叉, 突然変異操作) を行う。

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: 得られた最終世代の個体群の中から、式(2)の 2 目的において互いに非劣な個体を選択する。

3. 並列分散型 MoFGBML によるアンサンブル識別器

3.1. 並列分散実装

先行研究 [1] において、ファジィ GBML 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た部分集合に対して、複数の CPU コアを用いて MoFGBML を適用する。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。

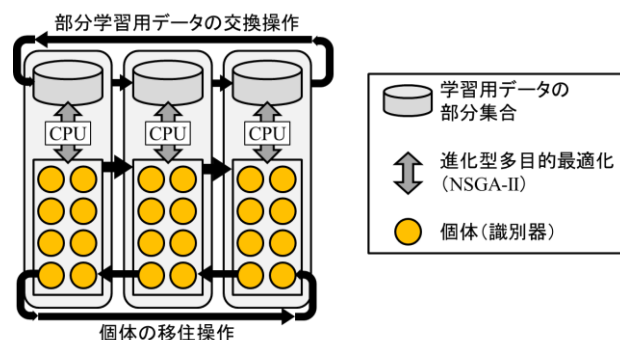


図 1: 並列分散実装のモデル図 [1]

3.2. 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単純多数決を行う。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識別器が存在しても、多数決による識別が可能となるため、識別性能の向上が期待される。また、単一の識別器と比較して、複数の弱識別器を用いることで多様性の高い識別器が設計できる。

本研究では、多様性の向上のため、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、多数決による識別精度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、多数決結果が同票多数となる可能性を低くできる。

4. 数値実験

4.1. 数値実験設定

本研究では, アンサンブル識別器の評価として, 交換操作, 移住操作を行った並列分散実装による結果と比較した. このとき, 交換間隔と移住間隔はともに 50 世代とした. これは, [4]において識別性能の良い結果を示したためである. 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 : 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

4.2. 数値実験結果

5. 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 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.
- [4] H. Ishibuchi, M. Yamane, and Y. Nojima, "Ensemble fuzzy rule-based classifier design by parallel distributed fuzzy GBML algorithms," *Proc. Of 9th International Conference on Simulated Evolution and Learning – SEAL 2012*, pp. 93-103, Hanoi, Vietnam, December 16-19, 2012