様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ９ 月　 ６ 日現在

機関番号：２４４０３
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００３３６

研究課題名（和文）多目的遺伝的機械学習手法による大規模多属性データからの知識獲得

研究課題名（英文）Data Mining from Large High-dimensional Data by Multiobjective Genetics-based Machine Learning

研究代表者
能島　裕介（Nojima, Yusuke）

大阪府立大学・工学（系）研究科（研究院）・准教授

研究者番号：１０３８２２３５

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：近年，データの利活用が盛んに行われ，同時に様々なデータマイニング手法の開発が行われている．我々が開発してきた多目的遺伝的機械学習は，複数のIf-thenルールで構成される知識を獲得するデータマイニング手法である．知識の精度最大化と複雑性最小化を目的関数として設定することで，手法の一度の実行で精度と複雑性の異なる様々な知識が獲得できるという利点がある．本研究では，大規模多属性データからの知識獲得をより効率的に行うために，アルゴリズムと実装面から多目的遺伝的機械学習手法の改良を行った．また，機械設計解析，画像データ識別問題，マルチラベル分類問題へと多目的遺伝的機械学習の拡張を行った．

研究成果の学術的意義や社会的意義
大規模多属性データからの知識獲得において，計算時間の短縮と得られた知識の分かりやすさは非常に重要である．本研究では，知識の精度と分かりやすさを同時に最適化する多目的遺伝的機械学習手法の高効率化を行った．また，これまでの数値データからの知識獲得だけでなく，様々な応用問題へと展開したことで，今後，解釈可能なAIの発展に寄与できると考えられる．

研究成果の概要（英文）：In recent years, data science is one of the hottest topics in computer science. Various data mining techniques have been developed so far. Among them, our multiobjective genetics-based machine learning is a data mining technique which can generate a number of knowledge models with different accuracy and complexity by its single run. In this research, we improved its effectiveness for large and high-dimensional data sets from both algorithmic and implementation points of view. We also extend our multiobjective genetics-based machine learning for the analysis on mechanical design problems, image data classification problems and multi-label classification problems.

研究分野：計算知能

キーワード：知識獲得　多目的最適化　並列分散実装

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年，データの利活用が盛んに行われ，それと同時に様々なデータマイニング手法の開発が行われている．多目的遺伝的機械学習は，複数の If-then ルールで構成される知識を獲得する進化計算に基づくデータマイニング手法である．知識の精度の最大化と複雑性の最小化を目的関数として設定することで，手法の一度の実行で精度と複雑性の異なるパレート最適な知識が獲得できるという利点がある．多目的遺伝的機械学習は，知識構造の様々な部分を利用者の要求に合わせて最適化することができる．例えば，条件部分の区間集合やファジィ集合の組合せ最適化や，それら集合の範囲をデータの分布に合わせて最適化することも可能である．また，多数のルール集合からのルール選択も可能である．このような利点から，多目的遺伝的機械学習，特にファジィルールを用いた多目的遺伝的知識獲得に関する研究が盛んに行われている．

多目的遺伝的機械学習は多くの利点があるものの，進化計算に基づく確率的多点探索手法であるため，より高精度な知識を得るためには多くの評価が必要であり，1 個体の評価に時間がかかる大規模データ（パターン数の多いデータ）に適用することが困難という問題がある．この問題に対して，荷重和重み付け目的関数を用いた単一目的遺伝的ファジィルール選択の並列分散実装やファジィ遺伝的機械学習の並列分散実装が提案されている．この並列分散実装は，島モデル型の進化計算のような個体群の分割だけでなく，学習用データも分割し，それぞれ別々の CPU コアに割り当てることで，CPU コア数の２乗倍の高速化が可能である．多目的ファジィ遺伝的機械学習に対しても並列分散実装を行い，計算時間の大幅な削減が可能であることは確認されたが，高精度の知識が獲得できないという問題が明らかになった．また，これまで並列分散実装で用いてきたデータ集合の規模（20,000 パターン以下）では，ビッグデータ解析には程遠い．さらに，1 パターンに含まれる属性数が多い多属性データへの対応も遅れているのが現状である．

２．研究の目的

多目的遺伝的機械学習は，進化計算を用いているため，１個体の評価に時間のかかる大規模データでは，知識を得るまでに膨大な計算時間を必要とする．ここでの大規模データとは，パターン数が多いデータと属性数が多いデータを意味する．属性数が多い場合，If-then ルールの条件部分の判定に時間がかかるため，結果的に個体の評価に時間がかかることになる．また，多目的遺伝的機械学習は，ベンチマークとして用いられる実世界データにのみ適用されてきたことから，実問題への応用に対する個別の対応が必要となる．

そこで本研究では，進化計算のアルゴリズム自体の改良，パターン数が多い大規模データに対する多数多数 CPU コアを用いた実装方法の開発，属性数が多い多属性データに対する効率的な探索方法の開発を行う．また，実世界問題への応用として，これまで扱ったことのない問題への適用を検討する．

３．研究の方法

３．１．多目的遺伝的機械学習の探索性能の向上

まず，ファジィ多目的遺伝的機械学習の並列分散実装 (Fig. 1) における探索能力，主に評価用データに対する汎化性能の向上に関して検証を行う．通常よりも多くの個体評価回数を用いることにより，非並列実装と並列分散実装を比較する．

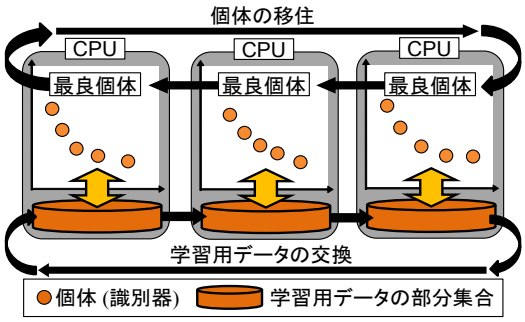

Fig. 1　並列分散実装 [2]

また，これまでの NSGA-II に加え，近年注目されている MOEA/D を用いた並列分散モデルの実装を行う．NSGA-II は，目的関数空間での解の優越関係により親個体選択と世代更新を行う．MOEA/D は，多目的最適化問題を目的関数空間に一様な複数のベクトルで単一目的最適化に分割し探索する手法である（Fig. 2 左図）．しかし，この手法では，典型的な実行可能領域に対して，端の解に対応するベクトルが疎であることから，誤識別率の小さな解の探索が弱い．そこで，探索方向を決めるベクトルの方向を修正した新しいスカラー化関数の提案を行う（Fig. 2 右図）．このベクトルの修正により，ルール数のように離散目的関数を効果的に分割することが可能となる．

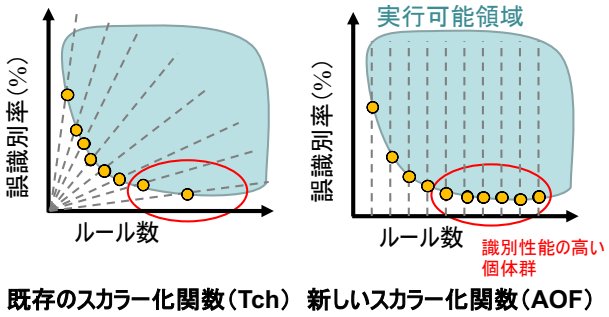

Fig. 2　既存のスカラー化関数と提案した新しい関数 [3]

３．２．多数 CPU コアを効率的に用いた実装

ビックデータからファジィ遺伝的機械学習で知識獲得を行う場合，並列分散実装は必須である．これまでの並列分散実装は，複数 CPU コアを持つ単一サーバ上で行い，個体群とデータ集合をそれぞれ使用する CPU コア数分だけ分割する方法であった．そのため，CPU コア数の多いサーバであれば，部分個体群サイズが小さくなりすぎ，探索能力が悪化することがあった．逆に，分割数を少なくする（部分個体群サイズをある程度大きく保つ）と，使用しない CPU コアが存在し効率が悪くなる．そこで，CPU コアの数と分割数を独立して扱う方法を提案する（Fig. 3 右図）[7]．提案実装は，複数のサーバを同時に用いることが可能であり，効率的に，計算時間の削減が可能であると同時に，我々の並列分散実装の特徴でもある分割個体群間での解候補の交換による探索性能の改善も可能である．

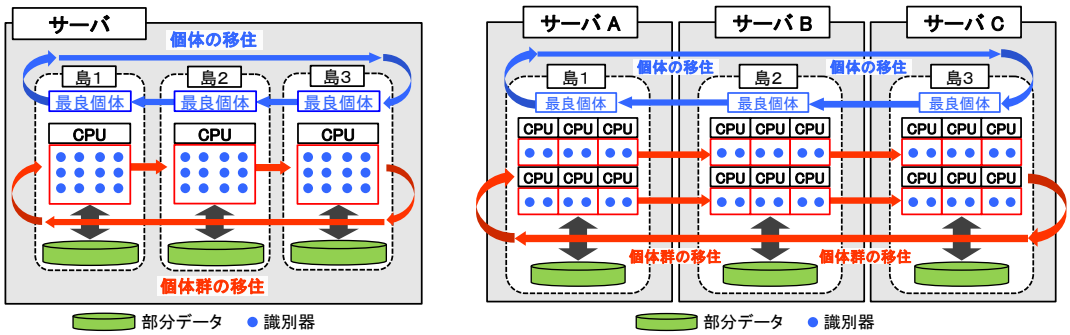

Fig. 3　複数 CPU による並列分散実装と複数サーバによる並列分散実装[7].

３．３．多属性データへの対応

多属性データに対して，誤識別パターンを中心とした複数パターンからのルール生成法の検討を行う [8]．属性数の大きなデータに関しては，多数のパターンからルールを生成することで汎化性の高い識別器が獲得可能であることを明らかにする．ルール生成に使用するパターン数をランダムに変化させる方法の提案と検討を行い，様々な個数のパターンに基づきルールを生成する方が良いことを明らかにする．

３．４．実世界問題への応用

実世界問題への応用として，多目的遺伝的機械学習を３つの問題に適用する．まず，多目的最適化問題解析への応用として，進化型多目的最適化により複数の解が得られた機械設計などの実問題から，設計変数と目的関数の関係を明確にする知識を獲得する手法の開発を行う [9], [10]. これまで評価した解の設計変数と目的関数値をパターンとして用い，特定の領域を複数選ぶことでクラス識別問題として定義する．さらに，多目的遺伝的機械学習を適用することで，If-then ルール集合の獲得を行う．

次に，画像識別への応用として，深層学習手法との併用方法を検討・開発を行う．通常の深層学習の出力では，単一のクラスラベルのみ推論されるが，提案手法では，複数のクラスラベルの尤度を入力情報とし出力を真のクラスラベルとした新たなパターン集合を定義し，そのパターン集合に対して，多目的遺伝的機械学習を適用することにより，識別器の設計を行う [11].

最後に，マルチラベル分類問題への多目的遺伝的機械学習の改良を行う [12]．これまで 1 つのパターンに 1 つのクラスラベルが付与されたデータを用いた識別器設計を行ってきたが，実世界には 1 つのパターンに複数のクラスラベルが付与されることも多い．そこで，複数クラスラベルに対応した多目的遺伝的機械学習手法の改良として，ファジィルールの後件部を拡張する．

４．研究成果

４．１．多目的遺伝的機械学習の探索性能の向上に関する実験結果

ファジィ多目的遺伝的機械学習の並列分散実装 [2] における探索能力および評価用データに対する汎化性能の向上に関して，通常よりも多くの個体評価回数を用いた比較実験結果を Fig. 4 に示す．

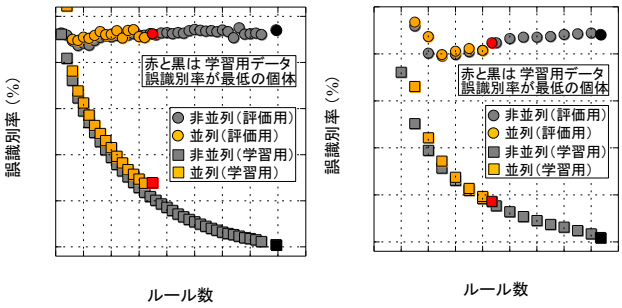

Fig. 4 比較結果．左：Cleveland data，右：Heart data

Fig. 4 に示す通り，非並列実装は個体評価回数を増やすことで，学習用データに対して精度の高い識別器が獲得できている．しかし，評価用データへの過剰適合も確認できる．一方，並列分散実装を行うことで，過剰適合が抑えられ，ルール数の少ない識別器が獲得できることを確認した[1], [2].

次に，最適化アルゴリズムに NSGA-II を用いた並列分散モデル (NSGA-II) と，チェビシェフスカラー化関数を用いた MOEA/D モデル (MOEA/D-Tch)，探索方向を決めるベクトルの方向を修正した新しいスカラー化関数 AOF（Fig. 2 右図）を用いたモデル (MOEA/D-AOF) の比較結果を Fig. 5 に示す．NSGA-II を灰色，MOEA/D-Tch を緑色，MOEA/D-AOF を青色で表す．逆三角形は，学習用データに対して最も誤識別率が低かった識別器を表す．Fig. 5 から分かるように，提案手法により，誤識別率の低い識別器が獲得できていることが分かる[3], [4].

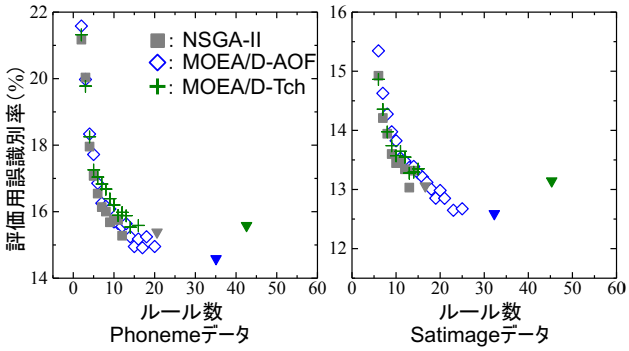

Fig. 5　異なる探索アルゴリズムの比較 [3]

さらに，ファジィ If-then ルール集合に基づく識別器設計だけでなく，区間集合を用いた識別器設計においても，並列分散実装やアンサンブル識別器の構築も行い，多目的遺伝的機械学習と並列分散実装の有効性を示した[5], [6].

４．２．多数 CPU コアを効率的に用いた実装に関する実験結果

まず，複数サーバを用いた知識獲得に対して，主流となりつつあるビックデータ解析ライブラリ(Spark)の利用を試みたが，並列分散多目的ファジィ遺伝的機械学習とはアルゴリズムの構造上相性が悪く，多少改良したところで大幅に計算時間が増加し，複数サーバを用いるメリットがないことが明らかになった．そこで無理にライブラリに合わせた改良は行わず，Java のソケット通信を利用したマスタスレイブ型の実装を行い，複数のサーバーを用いたスケールアップが可能であることを確認，1 千万パターンのデータまで実時間で適用可能であることを明らかにした．

Table 1 に実行時間の結果を示す．1 台 1 島モデルと 1 台 4 島モデルを比較すると約 3 倍の高速化が可能になったことが分かる．1 台 1 島モデルと 4 台 4 島モデルを比較すると約 6 から 8 倍程度高速化が可能になったことが分かる．また，Table 2 から，並列分散化により，汎化性能の高い識別器が獲得出来ていることも分かる．

４．３．多属性データに関する実験結果

多属性データに対して，誤識別パターンを中心とした複数パターンからのルール生成法を行った結果を Fig. 6 に示す．設定として 2, 5, 10, 20 個のパターンからルールを生成した場合，[2, 20]の範囲でランダムにパターン数を選択しルール生成を行った場合を比較した．また，世代更

新モデルとして(1, 1)-ES 型と(1+1)-ES 型を用意し比較を行った．

属性数の大きなデータに関しては，多数のパターンからルールを生成することで汎化性の高い識別器が獲得可能であることを明らかにした．また，ルール生成に使用するパターン数をランダムに変化させ様々な個数のパターンに基づきルールを生成する方が良いことを明らかにした．世代更新モデルとして(1+1)-ES 型を用いることで，複数パターンからのルール生成法の効果が大きくなることが分かった．

Table 1： 実行時間（秒）[7]

データ名	*1台1島*	*1台4島*	*4台4島*
Covtype	3,460	1,210	410
Poker	3,819	1,633	727
Susy	34,070	12,628	5,249
Hepmass	59,317	28,476	9,877
Higgs	77,628	26,736	11,839

Table 2： 評価用誤識別率（%）[7]

データ名	*1台1島*	*1台4島*	*4台4島*
Covtype	9.41	**9.34**	9.36
Poker	7.61	7.58	**7.57**
Susy	21.30	21.10	**21.05**
Hepmass	18.24	**18.11**	18.15
Higgs	36.17	35.44	**35.17**

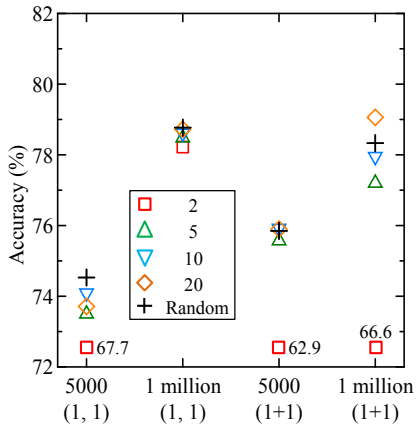

Fig. 6　7 つの多属性データに対する平均識別率 [8]

４．４．実世界問題への応用結果

多目的最適化問題解析に関して，機械設計問題における進化型多目的最適化手法により評価された解集合を用いて，設計変数の値を入力と目的関数値を出力とするパターン集合を生成した．そのパターン集合に対して，目的関数空間上で興味のある部分を複数選択し，クラスラベルを設定した．そのパターン集合に多目的遺伝的機械学習を適用することで，If-then ルールによる知識獲得を行い，注目している領域間の関係を少数の If-then ルールで表現できることを明らかにした [9]．また，進化型多目的最適化手法の実行中に上記知識獲得を行い，見込みのない解の評価を削減することで，パレート最適解の探索性能の改善が可能であることも明らかにした[10].

画像データ識別への応用に関して，複数のクラスラベルの尤度を入力とするパターン集合を作成し，多目的遺伝的機械学習を適用することで，既存のクラスラベルを用いた識別器の獲得を行った．獲得した識別器は，深層学習を用いた場合よりも識別率が改善され，さらに，クラスラベル間の関係性を明らかにできることを示した [11].

マルチラベル分類問題への応用に関して，複数のマルチラベルデータを用いた数値実験により，これまで提案されているマルチラベル分類手法と同等かそれ以上の性能が得らることを確認した [12]．多目的遺伝的機械学習を用いているため，高精度かつ複雑な知識から単純な知識まで得られることも示した．

５．主な発表論文等
〔雑誌論文〕（計 0 件）

〔学会発表〕（計 11 件）

[1] H. Ishibuchi, S. Takemura, and Y. Nojima, “Fitting and overfitting of multi-objective fuzzy genetics-based machine learning to training data,” *Proc. of 7th International Symposium on Computational Intelligence and Industrial Applications,* 6 pages, Beijing, China, November 3-6, 2016.
[2] 武村周治，能島裕介，石渕久生，多目的ファジィ遺伝的機械学習における並列分散実装の過学習に対する効果，第 10 回進化計算シンポジウム 2016 講演論文集，pp. 123-130，千葉，2016.
[3] 荒張巧樹，武村周治，能島裕介，石渕久生，多目的ファジィ遺伝的機械学習に特化したスカラー関数の提案，第 11 回進化計算学会研究会資料集, pp. 190-196，神戸, 2016.
[4] Y. Nojima, K. Arahari, S. Takemura, and H. Ishibuchi, “Multiobjective fuzzy genetics-based machine learning based on MOEA/D with its modifications,” *Proc. of 2017 IEEE International Conference on Fuzzy Systems*, 6 pages, Naples, Italy, July 9-12, 2017.
[5] Y. Nojima and H. Ishibuchi, “Effects of parallel distributed implementation on the search performance of Pittsburgh-style genetics-based machine learning algorithms,” *Proc. of 2016 IEEE Congress on Evolutionary Computation*, pp. 2193-2200, Vancouver, Canada, July 24-29, 2016
[6] H. Gao, Y. Nojima, and H. Ishibuchi, “Multi-objective GAssist with NSGA-II,” *Proc. of 18th International Symposium on Advanced Intelligent Systems*, pp. 696-703, Deagu, Republic of Korea, October 11-14, 2017.
[7] 武村周治，能島裕介，石渕久生，複数サーバを用いた並列分散型ファジィ遺伝的機械学習によるビッグデータ処理，第 13 回進化計算学会研究会講演集, pp. 106-109，滋賀, 2017.
[8] Y. Nojima, S. Takemura, K. Watanabe, and H. Ishibuchi, “Michigan-style fuzzy GBML with (1+1)-ES generation update and multi-pattern rule generation,” *Proc. of Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems*, 6 pages, Otsu, Japan, June 27-30, 2017.
[9] 能島裕介，谷垣勇輝，石渕久生，進化型多目的最適化により得られた解集合からの多目的知識獲得，第 10 回進化計算シンポジウム 2016 講演論文集，pp. 418-425，千葉，2016.
[10] Y. Nojima, Y. Tanigaki, N. Masuyama, and H. Ishibuchi, “Multiobjective evolutionary data mining for performance improvement of evolutionary multiobjective optimization,” *Proc. of 2018 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 741-746, Miyazaki, Japan, Oct. 7-10, 2018.
[11] Y. Nojima, S. Sakai, N. Masuyama, and H. Ishibuchi, “Multiobjective evolutionary classifier design using class scores by a deep convolutional neural network,” *PPSN 2018 Workshop on Evolutionary Machine Learning*, Coimbra, Portugal, Sep. 8-12, 2018.
[12] 荒張巧樹，増山直輝，能島裕介，石渕久生，マルチラベル分類に適応した多目的ファジィ遺伝的機械学習，第 12 回進化計算シンポジウム 2018 講演論文集，pp. 43-50，福岡，2018.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
なし

６．研究組織
(1)研究分担者
なし

(2)研究協力者
なし