様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年　5 月　27 日現在

機関番号：２４４０３
研究種目：若手研究(B)
研究期間：2010～2012
課題番号：22700239
研究課題名（和文）　並列分散遺伝的知識獲得における効率的な個体群およびデータの分割

研究課題名（英文）　Effective Population and Training Data Partitioning in Parallel Distributed Evolutionary Knowledge Acquisition

研究代表者
能島 裕介（Nojima Yusuke）
大阪府立大学・大学院工学研究科・助教
研究者番号：10382235

研究成果の概要（和文）：
ルール集合に基づく知識を獲得する方法として，進化計算を用いた遺伝的知識獲得手法が提案されている．この手法の問題点として，大規模データに適用した場合，膨大な計算コストが必要となる．本研究では，個体群と学習用データ集合を同時に分割し，複数の CPU を用いて計算する方法を提案し，計算時間の大幅な短縮と評価用データに対する汎化性の改善が可能であることを示した．また，個体群やデータ集合の分割方法の違いによる影響を調査した．

研究成果の概要（英文）：
Evolutionary knowledge acquisition has been proposed in order to obtain rule-based knowledge from numerical data. The main problem of this method is that huge computation cost is necessary when we apply it to large data. This study proposes parallel distributed implementation of evolutionary knowledge acquisition where both a population and training data are divided into subpopulations and training data subsets, respectively. The computational experiments show that the computational cost can be drastically reduced without the deterioration of the generalization ability. The effects of various specifications for parallel distributed implementation are also examined.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,800,000	540,000	2,340,000
2011 年度	700,000	210,000	910,000
2012 年度	600,000	180,000	780,000
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・感性情報学・ソフトコンピューティング
キーワード：遺伝的アルゴリズム

１．研究開始当初の背景

情報爆発時代において，顧客購買情報，医療情報，遺伝子情報など，大規模なデータから高精度かつ分かりやすい知識を獲得することが必要になってきている．まず，知識の分かりやすさに関して考えると，「もし○○であれば△△である」という If-then 形式の表現は直感的に受け入れやすいと思われる．このようなルール形式の知識獲得としては，相関ルールマイニングが有名であるが，個々のルールが表す事象はデータのほんの一部であり，全データを把握することはできない．そのため，ルール集合という形で知識を表現することが代替案としてあがられるが，膨大な数の相関ルールが獲得され，分かりやすさが失われる可能性がある．

この問題に対して，抽出した相関ルールの組合せ最適化を行う遺伝的ルール選択や，相関ルールマイニングを用いずにルール集合を直接最適化する遺伝的機械学習が提案されている．目的として精度の最大化と複雑性の最小化を行うことで，単純かつ高精度な知識を獲得することができるという利点がある．しかしながら，個体の評価に学習用データを繰り返し用いるため，大規模データに適用した場合，計算コストが大幅に増大するという問題がある．

２．研究の目的

本研究では，進化計算を用いた知識獲得手法である遺伝的ルール選択手法や遺伝的機械学習手法の高速化を目指し，並列分散実装を提案する．

進化計算の高速化で代表的な方法は，個体群を分割し，それぞれを別々の CPU で並列に進化計算を行う手法である．島モデル型とも言われ，分割数程度の高速化が可能である．部分個体群間での収束の回避と探索効率の改善を目的に，島の間で個体を交換する移住操作が適用される場合が多い．データマイニングの観点から高速化する方法として，データの一部を用いる方法である．パターンの削減や属性値の削減, Windowing などが代表的な方法である．

本研究では上記の２つのアイディアに基づいて，個体群の分割とデータ集合の分割を同時に行い，部分個体群の進化に部分データ集合を用いる方法を提案している．一対の部分個体群と部分データ集合を単一 CPU で処理することで，分割数の２乗倍の高速化が可能となる手法である．概略図を図１に示す．部分個体群が部分データ集合に収束しないように，一定世代ごとに部分データ集合と部分個体群の割り当ての変更を行う．また，探索性能を改善するために，部分個体群間での個体の移住操作も行う．本研究の目的は，この並列分散実装における構造の違いによる探索性能および計算効率の違いについて調査すること，および，各種パラメータ（部分データ集合の交換や移住間隔）の影響を調査することである．この並列分散実装の構造を生かしたアンサンブル識別器設計の構築や，クラスインバランスデータへの対応なども本研究の目的である．

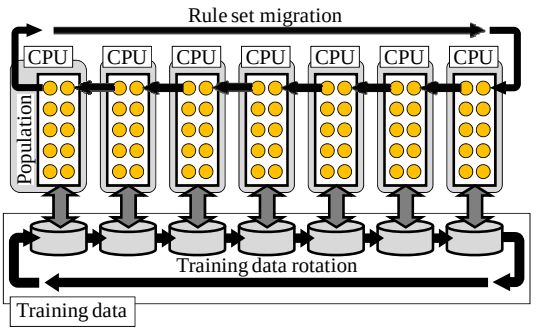

図１　並列分散型遺伝的知識獲得

３．研究の方法

(1) 遺伝的ルール選択

遺伝的ルール選択手法の並列分散実装において，部分データ集合の交換間隔の影響を調査する．条件部にファジィ集合を用いた場合，区間集合を用いた場合ともに，得られる知識（ルールに基づく識別器）の精度と計算時間に関して検討する．また，データ集合を CPU の個数以上に分割し，細分化した部分データ集合を用いた場合に関しても調査する．

さらに部分個体群における最良識別器をベース識別器とし，分割数分のベース識別器からなるアンサンブル識別器を設計する．

(2) ファジィ遺伝的機械学習

ファジィルールに基づく識別器の設計に遺伝的機械学習を用い，ルールの条件部の直接的な最適化を行う．この手法に並列分散実装を適用し，高速化を試みる．

探索性能の改善を目指し，エリート個体の移動（移住）を実装し，部分データ集合の移動間隔とともに，移住間隔の探索性能への影響を調査する．

さらに，データのみの分割，個体群のみの分割などとの比較を行い，提案している並列分散実装の構造的な利点を明らかにする．

遺伝的ルール選択と同様に，並列分散実装によるアンサンブル識別器の設計も行い，識別精度の改善を試みる．

(3) インバランスデータへの適用

特定のクラスのデータが，他のクラスよりも極端に少ないというインバランスデータからの知識獲得に本提案手法の適用を検討

する．具体的には，並列分散実装の構造的な特徴を利用し，多数派クラスのデータのみ部分データ集合へと分割し，少数派クラスに関してはすべての部分データ集合で同じものを用いることで，クラスバランスの偏りを緩和し，少数派クラスのデータも正しく識別できるような識別器の獲得を試みる．

(4) 進化型多目的最適化への拡張

精度の最大化と複雑性の最小化を同時に行い，精度と複雑性の異なるトレードオフを持つ複数の知識を一度に獲得する多目的知識獲得手法がある．この多目的知識獲得手法に対しても，並列分散実装を適用し，探索性能と計算時間の改善がどの程度行えるか調査する．

４．研究成果

(1) 遺伝的ルール選択

ルールの条件部に区間集合を用いた場合の結果の一部を以下に示す．Satimage データ（6435 パターン，36 属性，6 クラス）に対して，3 つの部分個体群と 3 つの CPU を用いた結果である．学習用データの分割を部分個体群数以上で分割し，異なる交換世代での影響を調査した．図 2 は，学習用データに対する識別率であるが，データの分割数が多くなれば識別性能が悪化することが確認できる．ただし，部分データ集合の交換を頻繁に行うことで，識別性能が改善していることが分かる．

図 3 は，評価用データに対する汎化性能である．黒く色づけされた結果は，通常の非並列非分散型の遺伝的ルール選択と統計的に有意な差がない設定を表している．部分データ集合を頻繁に交換することで，学習用データの分割数が 6 の場合でも汎化性能に差がない識別器が獲得できることが分かる．

図 4 に計算時間を示す．通常の非並列非分散型の場合，約 202 分かかった．データの分割が 6 の場合と比較すると約 20 倍の高速化が達成できていることが分かる．このように，汎化性能を落とさずに大幅な計算時間の短縮が可能であることを明らかにした．

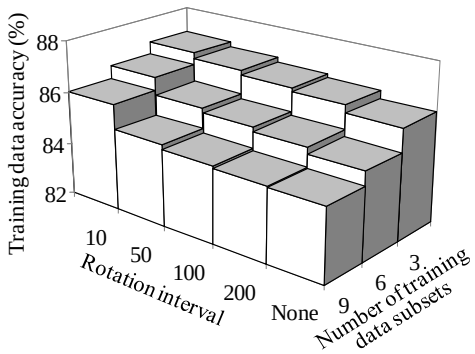

図 2　学習用データに対する識別率

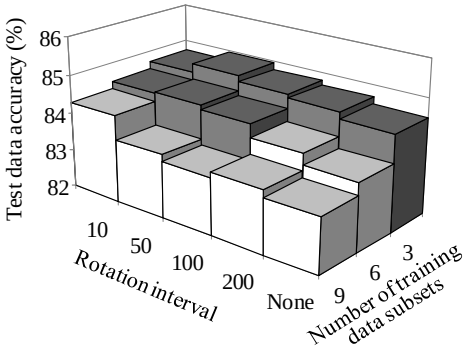

図 3　評価用データに対する識別率

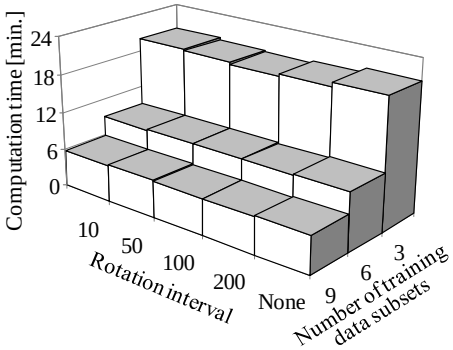

図 4　計算時間

(2) ファジィ遺伝的機械学習

ファジィ遺伝的機械学習手法の並列分散実装の結果を図 5 に示す．Ring データ（7400 パターン，20 属性，2 クラス）に対して，7 つの部分個体群と部分データ集合を用い，部分データ集合の交換間隔と移住操作間隔を変更した場合の誤識別率である．

部分データ集合の交換と移住操作の設定間隔によって，誤識別率が高い場合と低い場合が確認できる．誤識別率が高い結果は，部分データ集合を交換する際に，移住操作も同時に行った場合である．

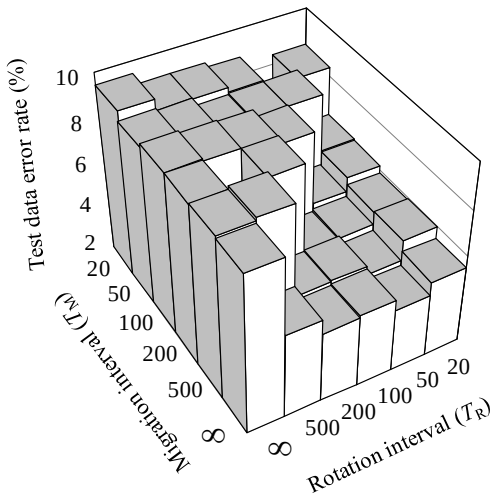

図 5　評価用データに対する識別率

この 2 つの操作を同時に行うことは，部分データ集合で最適化された個体がその部分データ集合の移動先にもコピーされることを意味する．これにより部分個体群の多様性が失われ探索が停滞することを明らかにした．

図 6 は，非並列非分散実装と並列分散実装の比較した結果を示す．比較のために，並列分散実装においても，全学習用データの誤識別率をモニターし，図 6 にまとめている．並列分散実装の結果から，誤識別率が世代に対して激しく上下していることが確認できる．これは，部分データ集合を交換したために，最適な解が入れ替わっていることを意味する．つまり，部分データ集合の交換により，部分個体群に摂動を加え，局所解への収束を防いでいると考えられる．

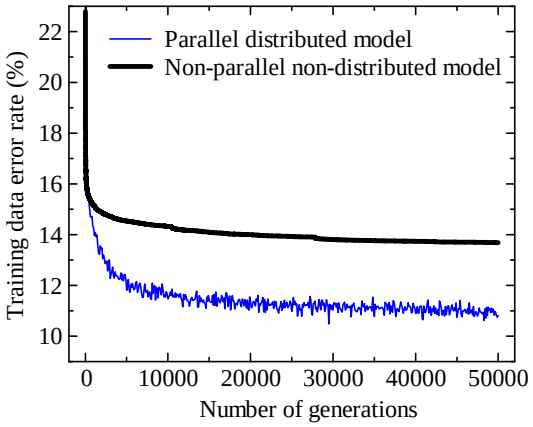

図 6　学習用データに対する誤識別率

(3) インバランスデータへの適用

図 7 に示すように，多数派クラスと少数派クラスのパターンがほぼ等しくなるように分割する方法を提案し，並列分散実装の枠組みで実験を行った．

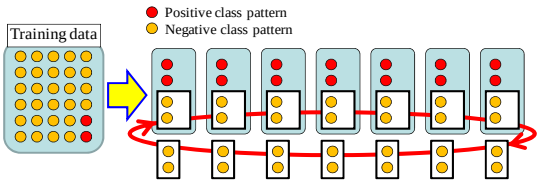

図 7　クラスインバランスデータに対するデータの分割方法

多数派と少数派クラスの割合が異なる複数のデータに対して，他手法との比較を行った結果を表 1 に示す．表 1 は平均ランキングを表しており，小さな値ほど良い．提案手法は，IR-based subdivision という名前にしているが，他手法と比較してもっとも良い平均ランクが得られた．

ファジィ遺伝的機械学習により，可読性の高い知識をインバランスデータからも獲得可能であることを明らかにした．

表 1　平均ランキング

Algorithm	Ranking
Simple subdivision	5.19
IR-based subdivision	**2.75**
Chi-3-LTR	4.13
Chi-5-LTR	5.63
HF-GBML-LTR	2.94
Ripper	4.56
C4.5	2.81

(4) 進化型多目的最適化への拡張

代表的な進化型多目的最適化手法として NSGA-II を用いて並列分散実装の多目的化を行った．並列分散実装による大幅な計算時間の短縮は可能であるが，探索性能が悪化するという結果が得られた．多目的最適化に適した個体群とデータ集合の分割方法の検討が今後の課題として残った．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① H. Ishibuchi, S. Mihara, and Y. Nojima, “Parallel distributed hybrid fuzzy GBML models with rule set migration and training data rotation,” *IEEE Transactions on Fuzzy Systems*, 査読有, vol. 21, no. 2, pp. 355-368, April 2013.

〔学会発表〕（計 11 件）

① 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*, 査読有, pp. 93-103, Hanoi, Vietnam, December 16-19, 2012.

② 山根優和，能島裕介，石渕久生，並列分散型ファジィ遺伝的機械学習によるアンサンブル識別器の設計，第 28 回ファジィシステムシンポジウム，pp. 715-718, 名古屋，9 月 12 日～14 日, 2012.

③ Y. Nojima, S. Mihara, and H. Ishibuchi, “Application of parallel distributed genetics-based machine learning to

imbalanced data sets," *Proc. of 2012 IEEE International Conference on Fuzzy Systems*, pp. 928-933, Brisbane, Australia, 査読有, June 10-15, 2012.

④ H. Ishibuchi, S. Mihara, and Y. Nojima, "Training data subdivision and periodical rotation in hybrid fuzzy genetics-based machine learning," *Proc. of 10th International Conference on Machine Learning and Applications*, 査読有, pp. 229-234, Honolulu, Hawaii, USA, December 18-21, 2011.

⑤ 三原新吾, 能島裕介, 石渕久生：並列分散型遺伝的機械学習における環境の変化が及ぼす汎化性能への影響, 進化計算シンポジウム 2011 講演論文集, pp.203-208, 宮城, 12 月 17 日～18 日, 2011

⑥ Y. Nojima, S. Mihara, and H. Ishibuchi, "Parallel distributed genetic rule selection of association rules," *Abstract Booklet of International Workshop on Simulation and Modeling related to Computational Science and Robotics Technology*, pp. 34-35, Kobe, Japan, November 1-3, 2011.

⑦ S. Mihara. Y. Nojima, and H. Ishibuchi, "Relation between migration interval and data rotation interval in parallel distributed fuzzy GBML," *Proc. of 12th International Symposium on Advanced Intelligent Systems*, 査読有, pp. 346-349, Suwon, Korea, September 29 - October 1, 2011.

⑧ 三原新吾, 能島裕介, 石渕久生：並列分散型ファジィ遺伝的機械学習の探索性能に対する個体移住操作の影響, 第 21 回インテリジェント・システム・シンポジウム講演論文集, (CD-ROM) 神戸, 9 月 12 日～14 日, 2011.

⑨ Y. Nojima, S. Mihara, and H. Ishibuchi, "Parallel distributed implementation of genetics-based machine learning for fuzzy classifier design," *Lecture Notes in Computer Science 6457: Simulated Evolution and Learning (8th International Conference on Simulated Evolution and Learning)*, 査読有, pp. 309-318, Springer, Berlin, December 1-4, 2010.

⑩ Y. Nojima, S. Mihara, and H. Ishibuchi, "Rotation effect of training data subsets in parallel distributed fuzzy genetics-based machine learning," *Proc. of 14th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, 査読有, pp. 96-105, Miyajima, Japan, November 19-20, 2010.

⑪ Y. Nojima, S. Mihara, and H. Ishibuchi, "Ensemble classifier design by parallel distributed implementation of genetic fuzzy rule selection for large data sets," *Proc. of 2010 IEEE Congress on Evolutionary Computation*, 査読有, pp. 2113-2120, Barcelona, Spain, July 18-23, 2010.

〔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〕
ホームページ等
http://www.cs.osakafu-u.ac.jp/~nojima/

６．研究組織
(1)研究代表者
能島　裕介 (Nojima Yusuke)
大阪府立大学・大学院工学研究科・助教
研究者番号：10382235

研究分担者　なし
連携研究者　なし