

オブジェクト指向を用いた 決定木と ADF GP の組み合わせによる学習

○新美 礼彦¹⁾ 田崎 栄一郎¹⁾
桐蔭横浜大学工学部制御システム工学科¹⁾

Combined Learning Method of Decision Tree and ADF GP by Object Oriented Approach

Ayahiko Niimi¹⁾Eiichiro Tazaki¹⁾

Department of Control & Systems Engineering, Toin University of Yokohama¹⁾

Abstract: There are many approaches for classification system learning. Genetic Programming (one of the approaches) can change trees dynamically, but its learning speed is slow. Decision Tree methods using C4.5 construct trees quickly, but the network may not classify correctly when the training data contains noise. For such problems, we proposed an object oriented approach, and a learning method that combines Decision Tree making method (C4.5) and Genetic Programming. To verify the validity of the proposed method, we developed a medical diagnostic system for the occurrence of hypertension, and compared the proposed method with prior methods.

Keywords: object oriented, genetic programming, decision tree, medical diagnostic system

1. はじめに

分類学習による推論システムの構築において、さまざまな手法が提案されている。一般に遺伝的プログラミングを用いた学習システムは、学習速度が遅く、システム設計者は問題と手法の両方の知識を要求される。しかし、構造も同時に扱うので、環境に適応したより高次の知識が獲得可能な学習システムの構築が可能である。また、決定木を用いた学習システムは、事例の分類モデルの構築に有効なネットワーク構造を得られ、他の手法と比較して短時間で学習を行えるが、トレーニング事例により分類精度が劣化するという問題がある。

このようにそれぞれの手法には利点と欠点がある。そこで本論文では、それぞれの手法をオブジェクトとして捉え、オブジェクトの組み合わせにより、学習を行う推論システムを構築するオブジェクト指向の手法を提案する。これにより、それぞれの手法の持つ利点と欠点を互いに補いながら学習が行われると期待される。

提案した学習法の有効性を検証するために、C4.5 による決定木構築法と自動関数定義を組み込んだ遺伝的プログラミングによる学習において、それらを組み合わせた学習を取り上げた。これを医療診断支援システムへ適用し、従来の単独での学習方法による結果と比較・検討し、ここに報告する。

手法の組み合わせの有効性は、決定木構築法と誤差逆伝播法によるニューラルネットワークの学習^[1]、決定木構築法と遺伝的プログラミングによるニューラルネットワークの学習^[2]によって検証されている。

2. 決定木構築のアルゴリズム^[3]

決定木構築のアルゴリズムに、記録された分類データを調べ、特定の例を一般化することによりモデルを帰納的に作る方法がある。決定木による分類学習は、比較的短時間である程度の分類能力を作ることができる。決定木構築法の 1 つである C4.5 では、期待獲得情報量最大化原理に基づく分類を行う。これにより、決定木の根に重要な属性を集めることができる。また、過剰に分類された決定木が構築されるのを防ぐため、予測誤り率による枝刈りを行う。

2.1 C4.5 のアルゴリズム

C4.5 による決定木構築は以下の手順に従う。

- 1) 初期決定木の構築
- 2) 構築された決定木に対する枝刈り

3. 遺伝的プログラミング

遺伝的プログラミング (Genetic Programming:GP) は、生物進化論の考えに基づいた学習法であり、そのアルゴリズムの流れは遺伝的アルゴリズム (Genetic Algorithm:GA) と同様である。その特徴は染色体表現が GA と異なり、構造表現ができるように拡張してあることである。今回は、決定木を表現するためにツリー構造を用いた。

3.1 GP のアルゴリズム

GP による決定木構築は以下の手順に従う。

- 1) 問題ごとの関数ノードと終端ノードのランダム文法から初期集団を発生させる。
- 2) 集団内にそれぞれの個体を計算し、問題解決にどのくらい関係しているかという適応度

(fitness value) を求める。

- 3) 遺伝的操作により、次の世代を発生させる。
 - a) 複製 (reproduction) により個体をコピーする。
 - b) 突然変異 (mutation) により新しい個体を発生させる。
 - c) 交叉 (crossover) により新しい個体を発生させる。
- 4) 終了条件が満たされたかどうかを調べ、満たされていたら終了する。満たされていなかった時は 2) へ戻る。

3.2 自動関数定義 [4]

通常、GP では木の大きさを評価しないため、木の成長を的確に制御する方法がない。このため、探索に伴って、木が長く複雑になることや逆に単純すぎる木に収束してしまうことがある。そこで関数をプログラム自身で定義して効率的に利用する方法が研究されている。そのうちのひとつが自動関数定義 (Automatically Defined Function: ADF) であり、これは通常の GP に関数定義用の遺伝子表現を付加することによって行われる。ADF を組み込むことにより、得られるプログラムがコンパクトにでき、また処理すべき個体も少なくできる。

4. オブジェクト指向

オブジェクト指向 (object oriented) とは、独立した情報処理を行う単位であるオブジェクトをプログラムの対象とみなし、オブジェクト間のメッセージのやりとりによって情報の処理を行う手法である。オブジェクトは入力情報を受信し、一連の処理を行い、出力情報を発信する。この際、自分自身の内部情報は隠している。このため、内部処理を気にすることなく、全体のデータの流れる考えられる。

5. 決定木と GP の組み合わせによる学習

決定木による分類学習は短時間の学習ですが、その訓練データにノイズが含まれていた場合、その分類能力は急速に劣化する。一方、GP は構造学習が可能であるため、高い分類能力を選られるが、学習の自由度が増すことにより、より多くの学習時間が必要となる。

この問題に対し、本論文ではこれら 2 つのアルゴリズムをオブジェクトと捉え、オブジェクト指向に基づくオブジェクトの組み合わせによる学習法を提案する。まず、C4.5 を用いて決定木を構築し、GP の初期集団に取り込み、学習を行う。これにより、GP の初期集団内に有効と思われるスキーマを含ませることが可能となり、学習速度と分類精度の改善が期待される。

6. 高血圧発症診断支援システムへの適用

我々が提案した学習法の有効性を検証するために、本論文では高血圧発症の実験データを用いて医療診断支援システムの構築を行った。このデータは、性別、年齢、肥満度など 15 項目からなる入力データと発症に関する出力の組み合わせで構成されており、全部で 1024 件ある。このうち、発症・未発症それぞれ 100 件ずつランダムに抽出し、こ

れを訓練データとして用いた。GP のパラメータは以下のものを用いた。(表 1 参照)

今回の GP では、定義されたノードの数が多く、生成される木の自由度が高かったため、決定木として使用できるものは、全個体中で極めてわずかであった。そのため、GP としての探索はランダム探索に近く推論精度の低下を招いたと思われる。初期個体として取り込まれた決定木は最良個体に継承されたことが確認され、学習効率の改善に影響を与えたといえる。

表 1 GP のパラメータ

集団数	500
複製確率	0.1
交叉確率	0.8
突然変異確率	0.1
選択方式	トーナメント方式
関数ノード	IFLTH, IFEQ, *, /, +, -, ADF0, ADF1
終端ノード	F00~F14, R, P, N
訓練データ数	発症(100), 未発症(100)
テストデータ数	1024

IFLTH, IFEQ: より少ない (<), 等しい (=)

ADF0, ADF1: ADF により拡張した関数定義遺伝子

F00 ~ F14: 入力データ

R: ランダムに生成される定数

P, N: 発症 (P), 未発症 (N)

表 2 各手法による実験結果 (推論精度)

	訓練データ	全データ	ノード数	獲得世代数
C4.5 のみ	98.5%	77.1%	29	-
ADF-GP のみ	75.0%	66.9%	148	14328
C4.5+ ADF-GP	96.0%	81.49%	17	41

7. おわりに

本研究では、オブジェクト指向を取り入れ、複数の手法を組み合わせる推論システムの構築を行うという学習手法を提案した。また、C4.5 と GP を取り上げ、単独で使った場合と組み合わせで使った場合について比較を行った。

その結果、わずかではあるが学習の効率の改善が認められた。このことより、提案した手法は推論システムの学習効率の改善に有効な方法であるといえる。

参考文献

- [1] A. Banerjee, R. Greiner(ed.), et. al.: Initializing Neural Networks Using Decision Trees: Computational Learning Theory and Natural Learning Systems, pp.3-15
- [2] 松本昇他: 決定木と組み合わせられた遺伝的プログラミングによるニューラルネットワークの創発的学習: 第 14 回ファジシステムシンポジウム, pp31-34
- [3] J. R. Quinlan, 古川康一 (訳), AI によるデータ解析: 凸版印刷株式会社, 1995
- [4] J. R. Koza, K. E. Kinner(ed.), et. al.: Scalable Learning in Genetic Programming Using Automatic Function Definition: Advances in Genetic Programming, pp99-117