相関ルールアルゴリズムと遺伝的プログラミングの組み合わせによる医療データの学習

新美 礼彦　田崎 栄一郎
桐蔭横浜大学 工学部 制御システム工学科

Combined Learning Method of Apriori Algorithm and Genetic Programming for Medical Database

Ayahiko Niimi　Eiichiro Tazaki
Department of Control and Systems Engineering, Toin University of Yokohama. Yokohama. Japan

Abstract: Genetic programming (GP) is characterized by its wide search space and a high flexibility, allowing GP to search for the global optimum solution. But, in general, GP's learning speed is slow. Apriori Algorithm is an association rule algorithm, and can be applied to large databases, but it is difficult to define its parameters without experience. We propose a technique for rule generation from database, using GP combined with the Apriori algorithm. This method takes rules generated by the association rule algorithm as initial population of GP. The learning speed of GP is improved by the combined algorithm. To verify the effectiveness of the proposed method, we apply it to the rule generation problem from medical database for meningitis. We compare the result of the proposed method with GP.

Keywords: Genetic Programming, Apriori Algorithm, Combined Learning, Data Mining, Medical Database

1. はじめに

医療データから診断支援システムを構築する手法は、数多く提案されている。遺伝的プログラミング(Genetic Programming:GP)は、システムの表現が比較的柔軟で、かつシステム構築時に構造学習を同時に行うことができる学習法である。これを医療診断支援システムの構築に利用できれば、簡単に高次の知識表現を含んだシステムの構築が可能であると考えられる。しかしながら、GPでは使用するノード数が増えると学習収束が遅くなったり、診断システムの精度の低下を招くことがある。

このような問題に対し、大規模な医療データからのルール構築をするために、相関ルールアルゴリズムとGPを組み合わせる手法を提案する。組み合わせて学習を行うことにより、より効率的な学習が行えるものと期待される。相関ルールアルゴリズムは、大規模データベースから高速でルール抽出を行うことができる。また、相関ルールアルゴリズムはGPにくらべて計算量が少ないので、組み合わせによる計算量の増加を防ぐことができる。

今回は、相関ルールアルゴリズムの中でも比較的高速なAprioriアルゴリズムを用いて実装を行った。これを医療データからルール生成の検証例として髄膜炎データベースに適用した。

2. 相関ルール分析

大量のデータから相関ルールを探索的に抽出するには、従来手法では多くの計算時間が必要であった。Aprioriアルゴリズムは、探索の効率を図ることにより大量データからでも現実的な時間で相関ルールを抽出することが可能である。[1)2)]Aprioriアルゴリズムでは、相関ルールの探索中に、支持度(support value)と確信度(confidence value)という2つの指標を用いて相関ルールの候補を評価する。

Aprioriアルゴリズムでは、支持度と確信度について最小支持度(minimun support value)と最小確信度(minimun confidence value)という閾値を設定し、各最小値を充たさない相関ルール候補をその相関性が低いものとして逐次的に評価対象から除外し、探索空間を縮小することにより、従来の相関ルール分析手法よりも計算時間について大幅に効率化している。

各最小値を操作することにより、対象とする相関ルール候補を増やしたり、探索範囲を縮小することができる。しかしながら、検索範囲を縮小すると意外なルールを抽出することができなくなる可能性がある。また、相関ルール候補が多すぎるとルールを分析する専門家の負担が増加し、有用なルールが埋もれてしまう可能性もある。

3. 遺伝的プログラミング

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。[3)]その特徴は染色体表現がGAと異なり、関数ノードと終端ノードを用い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GPを用いて決定木の学習を行う。そのため、各個体がそれぞれ決定木を表現できるように、関数ノードに条件文、終端ノードをそれぞれの属性

値とクラス名を用いた。また、本論文では、生成される決定木をコンパクトにするため、自動関数定義(Automatically Defined Function:ADF)を用いた。[4)]

4. 組み合わせによる学習

Aprioriアルゴリズムと GPの利点と欠点を補うため、本論文ではそれぞれの手法を組み合わせてルールの生成を行う学習手法を提案する。組み合わせ学習を行うことにより、大量のデータに対して、柔軟なルールを探索可能なシステムの構築が期待される。

組み合わせによる学習手法は以下の手順で行う。

1) Aprioriアルゴリズムを用いて相関ルールを生成する。
2) 生成された相関ルールをそれぞれ決定木に変換する。
   a. 相関ルールの条件部を決定木の分割属性としてとらえることにより、決定木の経路を構築する。
   b. この経路の終端ノードに相関ルールの結論部をおく。
   c. 相関ルールで定義されていない終端ノードに関しては、終端ノード候補の中からランダムに決定する。
3) 変換した決定木をGPの初期集団に取り込み、決定木の学習を行う。
4) GPの最良個体からルールを抽出する。

これにより、容易に相関ルールと決定木を組み合わせることが可能となる。また、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るので、GPの学習速度と分類精度の改善が期待される。

GPの最良個体からルールを生成するとき、GPの個体が決定木を表現しているため、Quinlanが提案している決定木からルールを作成する方法を利用した。[5)]それに加えて、GPより得られたルールは無効な条件や無意味な条件を多く含んでいることがあるので、これを取り除く処理を加えた。

5. 髄膜脳炎データによる検証実験

相関ルールを生成するアルゴリズムとして、Apriori アルゴリズムを用いた。AprioriアルゴリズムとGPを組み合わせて、ルール抽出を行うシステムを構築した。有効性を検証するために、髄膜脳炎データベースに対してルール抽出実験を行った。髄膜脳炎データベースは、データ数が140件と少ない。しかし、記述属性が34もあり、定義するノード数が一般的なGP学習で用いられる数よりも多いので、今回の検証に向いていると考えられる。実験の結果、GPのみを用いたものよりも、精度の良いルールを抽出することができた。ここで得られたルールの精度は、以前提案したC4.5とGPの組み合わせ学習[6)]を髄膜脳炎データベースに適用したときと同程度の精度であった。しかし、C4.5とGPの組み合わせの時は初期集団の多様性保持のために、決定木を複数作る必要があった。Aprioriアルゴリズムと GPの組み合わせでは、この問題を計算量をふやさずに改善できるため、今回提案する手法の利点はあると考えられる。

6. おわりに

本研究では、遺伝的プログラミングと相関ルールアルゴリズムの組み合わせにより、データベースからルールの生成を行う手法を提案した。また、提案した手法の有効性を検証するために、髄膜脳炎データを用いてルール生成を行い、その評価を行った。

その結果、ルールの精度の改善が認められた。このことより、提案した手法はルール生成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。また、相関ルールアルゴリズムとGPを組み合わせることにより、属性数が多くGPのみでは学習しにくい大規模データベースに対して、GP学習の可能性を開くことができたと考えられる。

今後は、他の医療データへの適用を検討するとともに、相関ルールから精度の高い決定木への変換アルゴリズムを検討する予定である。

参考文献

[1] 喜連川優、データマイニングにおける相関ルール抽出技法、人工知能学会誌、Vol.12 No.4,pp.513-520、1997.

[2] 寺邊正大, 片井修, 椹木哲夫, 鷲尾隆, 元田浩、相関ルールにもとづく属性生成手法、人工知能学会誌、Vol.15 No.1,pp.187-197、2000.

[3] 伊庭斉志、遺伝的プログラミング、東京電機大学出版局、1996.

[4] J. R. Koza, K. E. Kinner(ed.), et.al, Scalable Learning in Genetic Programming Using Automatic Function Definition, Advances in Genetic Programming, pp. 99-117, 1994.

[5] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman Publishers, 1993.

[6] 新美礼彦, 田崎栄一郎、決定木と遺伝的プログラミングの組み合わせによる知識構造発見手法、第38回人工知能基礎論研究会&第45回知識ベースシステム研究会、人工知能学会、pp.19-24、1999.