2F3-02

DT-ClGBI 法による肝炎データの解析

Analysis of Hepatitis Dataset by Using DT-ClGBI

茂木 明 Akira Mogi　Nguyen Phu Chien　大原 剛三 Kouzou Ohara　元田 浩 Hiroshi Motoda　鷲尾 隆 Takashi Washio

大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Osaka University

We analyzed the hepatitis data by Decision Tree Chunkingless Graph-Based Induction (DT-ClGBI), which constructs a decision tree for graph-structured data while simultaneously constructing attributes for classification. An attribute at each node in the decision tree is a discriminative pattern (subgraph) in the input graph, and extracted by Chunkingless Graph-Based Induction (Cl-GBI), which has been developed to overcome the problem that Graph-Based Induction cannot find overlapping patterns due to the nature of chunking without backtracking. To improve the predictive accuracy of the resulting decision tree, in this paper, we take an approach in which we first classify a given dataset into two classes, typical patients and untypical ones, and then apply DT-ClGBI to each class separately. Experimental results show that the proposed approach can improve the predictive accuracy and construct a more comprehensive decision tree than one resulting from applying DT-ClGBI to the whole dataset.

1. はじめに

肝生検は肝炎の進行程度を正確に計測できるが，検査費用が高く，また身体的負荷が大きいという課題がある．このため，血液検査や尿検査などの一般的な検査から肝炎の進行状況を予測することが重要となる．通常 1 回の検査で複数の項目について計測するが，1 回の検査における検査結果を 1 つのレコードに変換した場合，検査項目間にも病態を反映した相関があると考えられる．また，各レコードも患者の病態推移の影響を受けるため独立ではなく，時系列的な相関が強いと考えられる．両者の相関を同時に反映して肝炎の進行状況を予測するためには，同時期の検査値間の共起と時期の異なる検査値間の時系列的な共起を合わせて表現できるパターンを抽出し，抽出したパターンを用いて予測を行うことが重要となる．しかしながら，膨大，かつ複雑な時系列データから診断に有用なパターンを医師自身が見つけ出すことは極めて困難であるため，そのような知識発見を支援する系統的な手法の開発が急務となっている．

筆者らはこれまで，逐次ペア拡張（チャンキング）によりグラフ構造データから特徴的なパターンを抽出する Graph-Based Induction（GBI 法）[9, 3] を用いてグラフ構造データから属性，および属性値を生成しつつ，それらを利用して決定木を構築する Decision Tree–Graph-based Induction（DT-GBI 法）[1]，および GBI 法では同時に抽出できなかった部分的に重複するパターンを抽出可能な Chunkingless Graph-Based Induction（Cl-GBI 法）[2] を利用した決定木構築手法 Decision Tree Chunkingless Graph-Based Induction（DT-ClGBI 法）を提案し，千葉大学医学部附属病院からご提供頂いた肝炎データ [8] の解析を進めてきた [5, 4].

これまでの解析結果から，対象データにはある程度の偏りがあり，単純に DT-ClGBI 法を適用して構築された決定木ではどうしても正しく分類できない事例が存在し，そのような事例により全体の予測精度が下がっていることを確認した [4]. そこで本稿では，肝炎患者を典型的な患者と非典型的な患者の 2 つのクラスに事前に分類し，それぞれのクラスに対して DT-ClGBI 法を適用することで，得られる決定木の予測精度を改善する手法を提案する．ここでの典型的な患者とは，DT-ClGBI 法により構築した決定木に対して比較的予測精度が高い患者とする．評価実験では，肝臓の線維化の段階（程度）をクラスとし，血液検査結果の時系列のみで第 4 段階（肝硬変）の患者とそれ以外の段階の患者を分類する決定木を提案手法により構築した．その予測精度，および得られた決定木のサイズについて，対象データを事前に分類しなかった場合の結果と比較することで，提案手法の有効性を示す.

連絡先: 茂木　明
〒 567-0047　大阪府茨木市美穂ヶ丘 8-1
大阪大学産業科学研究所 元田研究室
電子メール：mogi@ar.sanken.osaka-u.ac.jp

2. DT-ClGBI 法

2.1 Cl-GBI 法のアルゴリズム

Cl-GBI 法では，グラフ中の接続された 2 つの隣接ノード（ノードペア）を数え上げ，上位 b（ビーム幅）個の頻出ノードペアを選ぶ．次に，それらの頻出ノードペアをチャンキングしてグラフを書き換えるのではなく，新たなノードラベルを割り当てることで擬似ノードとして扱う．グラフの書き換えをしないため，グラフ中に元から存在するノードは複数の異なる擬似ノードの構成要素となることが可能となり，重複パターンの抽出が可能となる.

このような Cl-GBI 法の基本手続きの概要を以下にまとめる．Cl-GBI 法の入力は，グラフ集合 D，ビーム幅 b，繰り返し回数 N_e，頻度の閾値 θ，ノードペア（パターン）の評価値の閾値 δ であり，出力は以下の手続きを N_e 回繰り返して抽出されたすべてのパターンのうち θ 以上の頻度をもつものの集合である．なお，パターンを評価する評価関数としては情報利得など頻度に基づくものが利用可能である．以下では，1 回の繰り返しをレベルと呼ぶ.

Step 1 グラフ中にある 2 つの連結されたノードからなるすべてのペアを抽出し，その頻度を数え上げる．ただし，レベル 2 以降では，少なくとも一方のノードが直前のレベルで登録された擬似ノードであるようなノードペアをのみを抽出する.

Step 2 それまでに抽出したノードペアのうちまだ擬似ノードとして登録されていないものの中から上位 b 個の頻出ペアを選択し，擬似ノードとして登録する．選択したペアを構

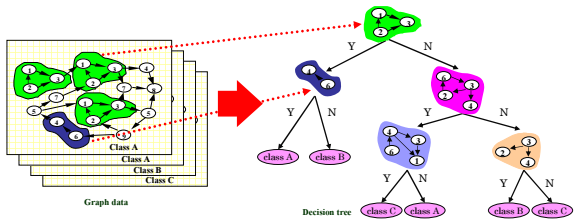

図 1: グラフ構造データを分類するための決定木

成するノードが擬似ノードであった場合，登録前に元のパターンに戻す．

Step 3 Step 2 で登録した擬似ノードに新たなノードラベルを割り当て，元のグラフは書き換えずに Step 1 へ戻る．

Cl-GBI 法では，探索が進むにつれて擬似ノードの数が増加し続け，それとともに考慮すべきノードペアの数も増加し続ける．ゆえに，探索範囲を規定するパラメータ b と N_e を適切に設定することがより重要となる．理論的には $\theta = 0$ としたときに b，N_e を十分に大きく設定することですべての部分グラフを見つけることが可能である．実際には，各レベル終了時に頻度が θ に満たないノードペアを抽出対象から除外することにより，無駄なノードペアの生成を回避することが可能となる．つまり，θ もまた探索空間を制御する重要なパラメータとなる．なお，Cl-GBI 法の詳細については参考文献 [2] を参照されたい．

2.2 DT-ClGBI 法のアルゴリズム

DT-ClGBI 法では，決定木の各分岐ノードにおいて上述の Cl-GBI 法により抽出した複数の特徴的なパターンを属性とし，各グラフにおけるパターンの有無を属性値とみなすことで分類対象となるグラフ集合に関する属性－属性値表を作成する．

次にその中から分類に効果的な属性（パターン）を選択する．属性値が“ yes（パターン有り）”と“ no （パターン無し）”のいずれかであるため，各分岐ノードでは選択されたパターンの有無に応じてグラフ集合が 2 分割される．このような操作を再帰的に繰り返すことにより，最終的に二分木として表現される決定木を構築する．

このようにして生成される決定木を図 1 に例示する．構築した決定木を用いてグラフとして表現された事例を分類する際には，事前に各分岐ノードに用いられたパターンを同様に Cl-GBI 法を使用して抽出しておく．また，DT-ClGBI 法のアルゴリズムを図 2 に要約する．パラメータ b，N_e，θ，δ は決定木の各ノードで独立に設定することができる．ただし，上位のノードで抽出されたノードペアが下位のノードで再度探索されることを回避するため，上位ノードで抽出したノードペアの情報はすべて下位ノードに引き継がれ，追加的に Cl-GBI 法でパターンを抽出する．また，図 2 のアルゴリズムにより構築された決定木には，過学習による予測精度への影響を軽減するために悲観的枝刈り [6] が適用される．

3. 肝炎データの解析

3.1 DT-ClGBI 法の 2 段階適用

筆者らはこれまでにも DT-ClGBI 法を用いた肝炎データの解析を進めており，その結果から対象データには DT-ClGBI 法で比較的精度よく分類可能な患者と，そうでない患者が存在することを確認している [4]．その結果，DT-ClGBI 法を用

DT-ClGBI（D）
Create a node DT for D
if termination condition reached
  return DT
else
  P := Cl-GBI（D） （with b, N_e, θ, δ specified）
  Select the best pair p from P
  Divide D into D_y （with p） and D_n （without p）
  for $D_i := D_y, D_n$
    DT_i := DT-ClGBI（D_i）
    Augment DT by attaching DT_i as its
    child along yes（no） branch
return DT

図 2: DT-ClGBI 法のアルゴリズム

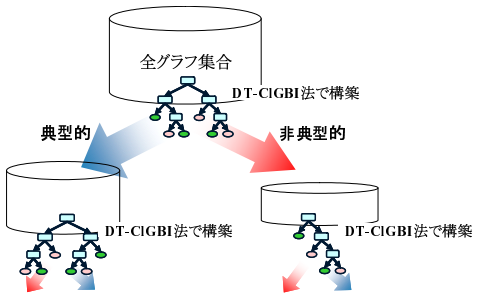

図 3: DT-ClGBI 法を 2 段階に適用して決定木を構築

いた 10-fold 交差検定により得られる決定木では，ルートノード，およびその直下のノードあたりでは概ね同じか類似した分岐パターンが得られるが，決定木の下部では安定しておらず，後者のいわゆる非典型的な患者の有無により予測精度が大きく影響を受けていた．

本稿ではこの問題を解消し，DT-ClGBI 法で得られる決定木の予測精度を向上させるために，図 3 に示すようにまず対象データを典型的，非典型的という 2 つのクラスに分類し，その各々に対して DT-ClGBI 法を適用するというアプローチを取る．具体的な手順は以下の通りである．

Step 1 対象グラフの集合に複数回 DT-ClGBI 法を適用する．

Step 2 各グラフの平均予測精度が事前に指定した閾値以上か否かで，対象グラフを典型的・非典型的の 2 つのクラスに分類する．

Step 3 典型的なグラフと非典型的なグラフを分類するための 1 段階目の決定木を DT-ClGBI 法で構築する．

Step 4 Step 1 における最良のパラメータを用いて，典型的・非典型的グラフ集合の各々に DT-ClGBI 法を適用し，本来のクラス分類のための決定木を構築する．

3.2 実験設定

肝炎データのうち肝硬変の患者（F4）と線維化が深刻でない患者（{F0+F1}）のデータに対して前節で提案した手法を適用した．それぞれのクラスラベルは，LC，nonLC とした．この場合，F4 の全 43 事例に対して {F0+F1} のクラスの事例総数が 129 事例と多いため，[7] と同様に各クラスの事例数の比が 2 : 3 となるように，LC の 43 事例に対して，F0 の全 4 事例と F1 から取り出した 61 事例の合計 65 事例を nonLC とした．実験では，[4] と同様に検査値の離散化，時系列デー

線維化の段階	F0	F1	F4
グラフ数	4	125	43
平均ノード数	303	304	300
最多ノード数	349	441	429
最少ノード数	254	152	162

表 1: 線維化の段階ごとのグラフサイズ

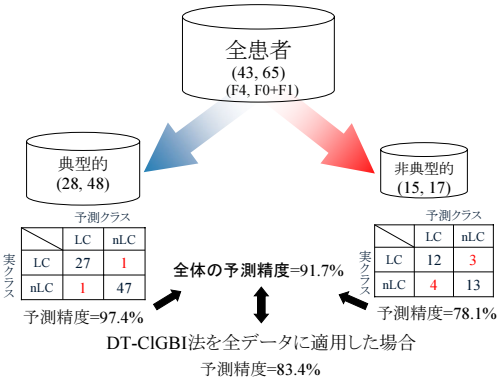

図 4: DT-ClGBI 法を 2 段階に適用して決定木を構築した結果

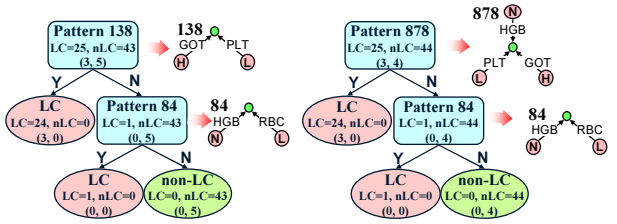

図 5: 2 段階目で構築された決定木例 (典型的グラフ集合)

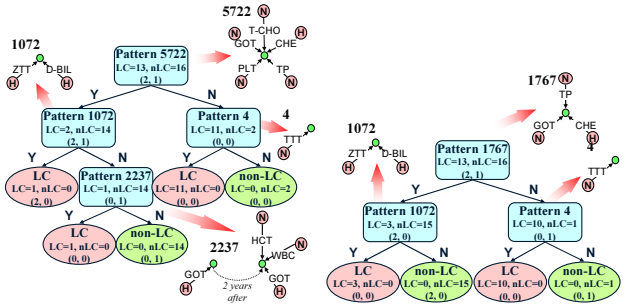

図 6: 2 段階目で構築された決定木例 (非典型的グラフ集合)

タの平均化等の前処理の後，1 人の患者のデータを 1 つのグラフに変換して用いた．変換後のグラフサイズを表 1 に示す．

提案手法の Step 1 に関しては，DT-ClGBI 法を用いた 10-fold 交差検定を複数のパラメータ設定で実施した．具体的には，評価値（情報利得）の閾値 δ を 0.01，頻度の閾値 θ を 0.1 に固定し，パラメータ b, Ne をそれぞれ $b = \{5, 6, 8, 10\}$, $Ne = \{6, 8, 10, 12\}$ の範囲で変化させた．すなわち，全 16 通りのパラメータ設定で 10-fold 交差検定を実施した．なお，今回の実験においては計算時間の観点から，決定木のルートノードでのみ Cl-GBI 法を実行し，その際に抽出されたパターンを下位の分岐ノードでの分類にも用いた．これは，ルートノード以外の分岐ノードでは Cl-GBI 法のパラメータ b, Ne をそれぞれ $b = 0$，$Ne = 0$ としたことを意味する．Step 4 における 2 段階目の決定木構築に関しては，Step 1 における最良パラメータ（$b = 8$, $Ne = 10$）を用いて，DT-ClGBI 法を適用した．

また，本実験では上記手順のうち Step 3 における 1 段階目の決定木は構築しなかった．これは，LC と nonLC に関する予測精度をこれまでの結果と純粋に比較するためである．ゆえに，本実験では上記手順の Step 4 において典型的・非典型的グラフ集合それぞれに対して DT-ClGBI 法で決定木を構築する際に 10-fold 交差検定を実行し，その予測精度を求めた．なお，Step 2 における典型的・非典型的なクラスへの分類については，平均予測精度の閾値を 100%とした．

3.3 結果と考察

実験結果の概要を図 4 に示す．Step 2 では LC に属する 43 個のグラフと nonLC に属する 65 個のグラフのうち，典型的なクラスに分類されたのはそれぞれ 28 個（LC）と 48 個（nonLC）であり，非典型的なクラスに分類されたのはそれぞれ 15 個（LC）と 17 個（nonLC）であった．その結果，典型的な患者に関する予測精度は 97.4%，非典型的な患者に関する予測精度は 78.1% となり，全体の予測精度は 91.7%となった．DT-ClGBI 法を全データにそのまま適用した 10-fold 交差検定で得られていた予測精度が 83.4%（$b = 8, N_e = 10$）であったことを考えると [4]，提案手法はその予測精度を大幅に改善したといえる．ただし，非典型的な患者に関する予測精度はそれほど高くないことから，そのような患者に対しては予測精度を改善する何かしらの対応が必要である．

Step 4 で典型的な患者，および非典型的な患者それぞれのクラスに対して構築された決定木の例を図 5，および図 6 に示す．図からもわかるように，いずれのクラスにおいても，決定木の上位のノードに現れる分岐パターンは同一か非常に類似していた．また，決定木のサイズに関しても，典型的な患者に対する決定木の平均サイズは 4.8，非典型的な患者の場合は 9 であり，DT-ClGBI 法を全データに適用した場合の平均サイズである 14.2 と比較して，大幅に小さくなっていることが分かる．これらの結果から，事前に事例を分類することによりその特徴が集約され，より簡潔で解釈しやすい決定木が得られたと考えられる．個々のパターンをみても，「GOT の値が高い」，「血小板（PLT）の値が低い」など肝炎に特徴的な傾向が上位のノードに見られており，なんらかの専門的な意味を持つのではないかと推測できる．その妥当性については，今後，専門家（医師）による評価を受ける予定である．

4. おわりに

本稿では，肝炎データ解析において事前に事例を典型的な患者と非典型的な患者の 2 クラスに分類し，それぞれのクラスに DT-ClGBI 法を適用する手法を提案し，実験的にその有効性を評価した．評価実験を通して，事前に事例を分類することで，DT-ClGBI 法を単純に全データに適用する従来手法と比較して，予測精度が約 10%ほど向上し，より簡潔で解釈しやすい決定木が得られる事を確認した．

今後は，非典型的な患者に関する予測精度の改善を図るとともに，典型的な患者と非典型的な患者を分類する 1 段階目の決定木を構築した場合の評価をする必要がある．加えて，提案手法で非典型的と分類した患者と実際の例外的な患者の関係についても検討したい．

参考文献

[1] Geamsakul, W., Matsuda, T., Yoshida, T., Motoda, H. and and Washio, T.: Performance evaluation of decision tree graph-based induction, *Proc. of the 6th Pacific-Asia Conference on Discovery Science (Springer Verlag LNAI2843)*, pp. 128–140 (2003).

[2] Nguyen, P., Ohara, K., Motoda, H. and Washio, T.: ClGBI: A novel strategy to extract typical patterns from graph data, *SIG-KBS-A403*, pp.105–110(2004).

[3] 松田, 元田, 鷲尾: 一般グラフ構造データに対する Graph-Based Induction とその応用, 人工知能学会誌, Vol.16, No.4, pp.363–374(2001).

[4] 茂木, Nguyen, 大原, 元田, 鷲尾: DT-ClGBI 法による肝炎データからの知識発見, 人工知能学会研究会資料, *SIG-KBS-A405*, pp.19–25(2005).

[5] Ohara, K., Yoshida, T., Geamsakul, W., Motoda, H., Washio, T., Yokoi, H. and Takabayashi, K.: Analysis of Hepatitis Dataset by Decision Tree Graph-Based Induction, *Proc. of Discovery Challenge, Workshop held in conjunction with the 8th PKDD*, pp. 173–184 (2004).

[6] Quinlan, J.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers(1993).

[7] Yamada, Y., Suzuki, E., Yokoi, H. and Takabayashi, K.: Decision-tree induction from time-series data based on a standard-example split test, *Proc. of the 12th International Conference on Machine Learning*, pp.840–847(2003).

[8] 山口：慢性肝炎データセットのクレンジングとマイニングの試み，平成 13 年度科学研究費補助金 特定領域 (B) 研究成果報告書，情報洪水時代におけるアクティブマイニングの実現, pp. 205–221(2002).

[9] 吉田, 元田: 逐次ペア拡張に基づく帰納推論, 人工知能学会誌, Vol.12, No.1, pp.58–97(1997).