2F2-01

Decision Tree – Graph-Based Induction の探索能力改善

Improvement of Search Capability of Decision Tree – Graph-Based Induction

ワロドム・ジアムサクン[*1] 松田 喬[*1] 吉田哲也[*1] 元田 浩[*1] 鷲尾 隆[*1]
Warodom Geamsakul Takashi Matsuda Tetsuya Yoshida Hiroshi Motoda Takashi Washio

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

Abstract: Decision Tree – Graph-Based Induction (DT-GBI) is a technique of constructing data classifiers (decision trees) from graph-structured data in which attributes and attribute values are not explicitly expressed. In our approach, attributes, namely substructures useful for classification task, are constructed by GBI on the fly while constructing a decision tree. After we proposed DT-GBI, the improved version called Beam-wise GBI (B-GBI) in which the idea of beam search was incorporated was proposed by Matsuda et al. in 2002. In this paper, we improved our DT-GBI by introducing B-GBI for typical substructure extraction. We used a DNA sequence dataset from UCI machine learning repository to evaluate the predictive accuracy of our technique with conventional GBI and new B-GBI. The results show that search capability of DT-GBI was improved by expanding search space when extracting typical substructures from graph-structured data.

1. はじめに

データからの知識発見ではグラフを扱えなければ解くことができない問題が多数存在し、その一例として化学物質が挙げられる。物質の構造からその性質などの予測を試みることができれば、物質を合成する前にその有害性・安全性を事前に評価でき、新規化学物質の開発にとって意義は極めて高い。また、ネットサーフィン状況や記号配列などもグラフ構造として表現することができる。このように、グラフとして自然に表現できるようなデータから知識を発見するにはグラフ構造データからのマイニングが望まれる。

有向グラフから類型パターンを発見する一つの手法として Graph-Based Induction（GBI 法）があり、GBI 法ではペアを逐次拡張していくことを基本原理とする。評価関数は主として統計的な指標を用い、選ばれたペアを逐次的に拡張することを通じて有向グラフで表されたデータからの概念の抽出を実現してきた [吉田 97, 松田 01]。また、探索空間を広げるためにビーム探索を取り入れた Beam-wise Graph-Based Induction（B-GBI 法）も後に提案された [松田 02]。

他方、決定木 (decision tree) は、分類や予測を行う際に広く使われる手法である。利点として、決定木より導出される「ルール」が理解しやすいことが挙げられる。しかし、決定木を構築するためには、属性及び属性値が明示的に表現されている必要がある。クラフ構造データでは、属性を事前に定義できないため、グラフ構造データの分類に対して直接決定木として表現される分類木を構築することは困難であった。

著者らは従来の GBI 法でグラフ構造データから属性および属性値を生成し、決定木を構築していく手法 Decision Tree – Graph-Based Induction（DT-GBI 法）を提案してきた [ジアムサクン 02]。本稿では、GBI 法の探索方法を拡張した B-GBI 法を用いて属性および属性値を生成し、DT-GBI 法の予測精度が向上することを報告する。さらに、UCI Machine Learning Repository [Blake 98] における DNA の塩基列データセットを用いた DT-GBI 法の評価結果を報告する。

連絡先: ワロドム・ジアムサクン
〒 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：warodom@ar.sanken.osaka-u.ac.jp

2. Beam-wise Graph-Based Induction（B-GBI 法）

2.1 従来の GBI 法

GBI 法はグラフ構造データ中に現れる特徴的なパターンを抽出することを目的に考案された [吉田 97]。GBI 法は図 1 に示すように「ペアの逐次抽出（チャンキング）により特徴的なパターンを抽出する」という基本的な考えにより実現されている。ここで、「ペア」とは「二つのノードおよびそれらをつなぐリンク」からなる GBI 法で用いる基本単位となるものである。また、ペアは逐次拡張される（チャンクされる）ことにより複雑なパターンになっていく。

GBI 法では頻度に基づく任意の評価関数を使用できる。但し、逐次ペア拡張に基づくため、特徴的なパターンを見つけるためにはその部分パターンも特徴的なものでないといけないという条件がある。図 1 では、1→3 ペアが特徴的でチャンクされない限り、2→10 ペアが見つかることはない。特徴的なパターンを頻度で評価することにすれば、頻度はこの単調性を満たす。評価関数がこの条件を満たさなければ、その評価値が最も高いペアを選んでチャンキングを数回繰り返しても良いパターンが得られない可能性がある。この問題を解決するために、松田らは 2 つの評価関数を同時に使うように GBI 法を改良した。評価関数の一つはチャンクするペアを選択するための頻度、もう一つはグラフ中に含まれるペアの中から特徴的な部分グラフを抽出するための任意の関数である。なお、二つ目の評価関数は単調性を満たさなくても良く、情報利得 [Quinlan 86] や、利得比 [Quinlan 93]、ジニ指数 [Breiman 84] などを用いることができる。

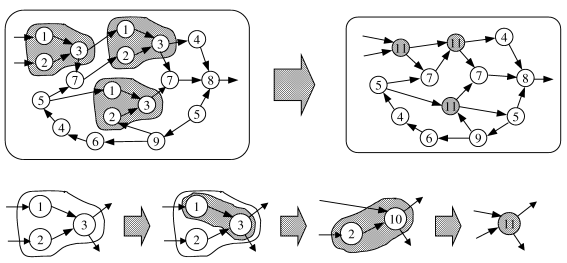

図 1: 逐次ペア拡張の基本的な考え

2.2 ビーム探索

グラフから全ての部分グラフを抽出することは NP 完全問題であるため、GBI 法は全ての特徴的パターンを抽出のではなく、ある程度の大きさを持った有意な部分グラフのみ抽出する。そのため、大規模なグラフ構造データから特徴的なパターンを抽出するには有用である。しかし、欲張り探索のために、評価値が同値のペアが複数存在した場合や、同じノードラベルが多数存在している場合には同種類のペアの連鎖が生じるためにチャンクすべきペアの選択に曖昧性が生じる。また、評価値がもっとも高い部分グラフしか選ばないために、のちに現れる可能性があるもっと特徴的な別のパターンを見逃してしまうことが十分考えられる。

GBI 法にビーム探索を導入することで、この問題が低減される。具体的には、チャンクするペアを唯一つではなく、ある一定の数だけ選択してそれぞれのペアについてチャンクする。これにより、複数の並列な状態に分裂する。次のステップではそれぞれの状態についてさらにある一定の数だけのチャンクすべきペアを選択するのではなく、全ての状態からパターンを一定の数だけ選択してチャンク候補とする。これにより、チャンクが進んでいくにつれて状態の数が爆発的に増加することを防ぐことができる。

2.3 B-GBI 法のアルゴリズム

B-GBI 法が行う「逐次ペア拡張」アルゴリズムは以下の通りである。この作業が終了条件（通常は最低サポート）が満たされるまで繰り返される。

ステップ 1 全ての状態について、グラフに存在するペアを全て抽出する。

ステップ 2a ステップ 1 で抽出したペアのうち、評価関数により特徴的なペアを全て登録する。この時、ペアを構成するノードが既に書き換えられたノードであれば元のパターンに復元してから登録する。

ステップ 2b ステップ 1 で抽出したペアのうち、頻度によりチャンクすべきペアをある一定の数だけ選び、抽出パターンとして登録する。この時、ペアを構成するノードが既に書き換えられたノードであれば元のパターンに復元してから登録する。この際，チャンクすべきペアがなければ終了する。

ステップ 3 ステップ 2b で選ばれたそれぞれのペアに対し、ペアを一つのノードに置き換えることにより、それぞれにグラフを書き換える。この際、必要の応じて状態を分裂または消滅させる。そして、ステップ 1 に戻る。

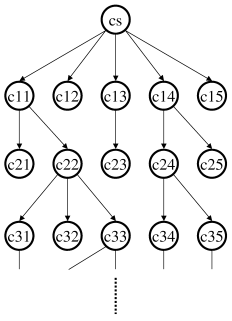

図 2: GBI 法におけるビーム探索（ビーム幅=5 の場合）

3. Decision Tree – Graph-Based Induction（DT-GBI 法）

3.1 B-GBI 法での要素の構築

決定木の表現は理解しやすいため、属性 - 属性値で表されるデータに対する分類器としてしばしば利用される。一方、グラフ構造データは通常ノードとリンクで表現され，属性と属性値に相当する要素を持たない。このため、グラフ構造データに対して直接決定木を構築することは困難である。しかし、部分グラフを属性とすれば、グラフ構造データを属性 - 属性値として表現できるため、決定木を構築することが可能になる。

分類に有効な部分グラフを予め抽出することは困難であるが、ペアを B-GBI 法により逐次的に拡張し、特徴的なペアを決定木を構築しながらに拡張していくことで、決定木の構築と同時に分類のための属性に相当する特徴的なパターン（部分グラフ）を作ることができる。著者らの手法では、属性および属性値を次のように定義する。

- 属性：グラフ構造データに含まれるペア（部分グラフ）
- 属性値：グラフ中でのペア（部分グラフ）の有無

決定木を構築する時、全グラフに含まれる全てのペアを数え上げ、その中から評価値の高い（分類に効果的と考えられる）ペアを一つ選ぶ。データ（構造データの集合）を二つのグループ、つまり、選ばれたペアが含まれるグループと含まれないグループに分ける。次に、前者に属する全てのグラフに対して、選ばれたペアを一つのノードに置き換えるチャンキングを行う。選ばれたペアが 1 枚のグラフに複数含まれる場合は全てチャンクする。これらの過程を決定木の各ノードで実行し、分類のための属性を作ると同時に決定木を成長させていく。以上の DT-GBI 法のアルゴリズムを図 3 にまとめる。なお、属性値が YES（指定のペアがある）と NO（指定のペアがない）の二つであることから、構築された決定木は二分木になる。

以上で提案した手法は、決定木を構築しながら分類のための属性（ペア）を構築するという特徴を持つ。データ分類のためのペアが選ばれる度にそのペアがチャンクされ、より大きな部分グラフに成長していく。従って、初期のペアに二つのノードとそのリンクしかなくても、数回チャンキングを適用することで分類に効果的なペアが徐々により大きなペア（部分グラフ）に成長する。提案した DT-GBI 法では分類に有効な属性（ペ

DT-GBI(D)
  Create a node DT for D
  if termination condition is reached
    return DT
  else
    P := B-GBI(D) (with the number of chunking specified)
    Select a pair p from P
    Divide D into D_y (with p) and D_n (without p)
    Chunk the pair p into one node c
    D_{yc} := contracted data of D_y
    for D_i := D_{yc}, D_n
      DT_i := DT-GBI(D_i)
      Augment DT by attaching DT_i as its child along yes(no) branch
  return DT

図 3: DT-GBI 法のアルゴリズム

表 1: 第 1 段階での属性 - 属性値表

Graph	a→a	a→b	a→c	a→d	b→a	b→b	b→c	b→d	c→b	c→c	d→a	d→b	d→c
1 (class A)	1	1	0	1	0	0	0	1	0	0	0	0	1
2 (class B)	1	1	1	0	0	0	0	0	0	1	1	1	0
3 (class A)	1	0	0	1	1	1	0	0	0	0	0	1	0
4 (class C)	0	1	0	0	0	1	1	0	1	0	0	0	0

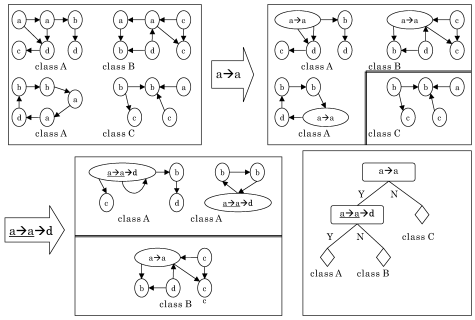

図 4: DT-GBI 法の適用例

ア）が次々と構築されることから、属性構築の手法の一つと考えることができる。

3.2 DT-GBI 法の適用例

DT-GBI 法の適用例を図 4 に示す。この手法が図 4 の左上隅のようなグラフ集合を入力とする。このデータには次の 13 種類のペアがある；a→a、a→b、a→c、a→d、b→a、b→b、b→c、b→d、c→b、c→c、d→a、d→b、d→c。a→a ペアがクラス A とクラス B のグラフに存在し、クラス C のグラフに存在しない。この段階でのペアの有無を表 1 に示すような属性 - 属性値表に変換する。その後、最も評価値の高いペアはデータを 2 つのグループ（選ばれたペアが含まれるグループと含まれないグループ）に分けるものとして選ばれる。ペアが含まれるグループで現れる全ての分岐ペア（選ばれるペアそのもの）が一つのノードにチャンクされ、グラフが書き換えられる。以上の過程を各決定木ノードで再帰的に繰り返す。最後に、それぞれのグループ内で同一クラスのデータしか入っていない状態に達するとデータの分岐が終了し、図 4 の右下に示す決定木が得られる。

4. DT-GBI 法に対する評価実験

著者らは前節で示したアルゴリズムを実装し、評価実験として DNA の塩基列データに適用し、クラス分類に適したパターンを抽出した。用いたデータセットは UCI Machine Learning Repository [Blake 98] により提供されている Promoter データセットである。Promoter データセットは塩基を表す A（アデニン）、T（チミン）、C（シトシン）、G（グアニン）からなる長さ 57 の文字列データであり、クラスはその塩基列が “Promoter”（DNA を鋳型に mRNA 合成を開始する DNA 上の特定の塩基列）を含むことを示す Promoter と “Promoter” を含まないことを示す Non-promoter の 2 つである。データセット中の事例数は 106 個で、クラス Promoter とクラス Non-promoter のデータがそれぞれ 53 個である。

このデータセットの説明および分析の詳細は文献 [Towell 93] を参照されたい。このデータはある位置を基準に塩基が整列されるように準備されたものであり、属性 - 属性値表現に従って n 番目の属性に n 番目の塩基を割り当てることができる。松田らの実験 [Matsuda 02] では、C4.5 [Quinlan 93] を用いた Leave-One-Out (LVO) による評価実験では予測誤り率が

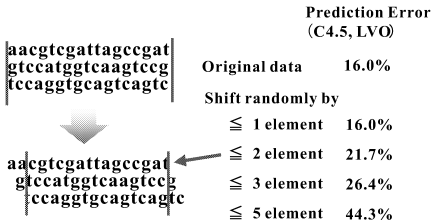

図 5: 誤った予測

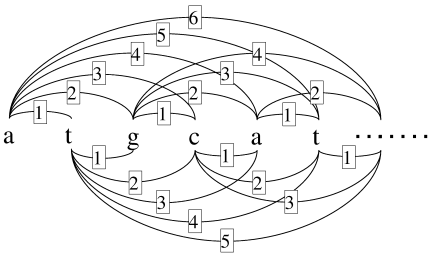

図 6: DNA 配列からグラフへの変換

16.0%であり、ランダムに配列を 5 つシフトすると予測誤り率が 44.3%となった（図 5 参照）。このことは、データが正しく整列されなければ、C4.5 など属性 - 属性値表現を使う標準的な分類器の予測精度が悪くなることを示している。

グラフ表現の利点の一つとして、特定の位置を基準にデータを整列しなくても良いことが挙げられる。本稿では、この文字列を図 6 のように塩基をノードラベルとし、一つの塩基から両側 1〜10 個の塩基にそれぞれ 1 から 10 のラベルをつけたリンクを張ることで一つのグラフに変換し、B-GBI 法への入力とした。一つのグラフのサイズはノード数 57 個、リンク数 515 本と、かなり大きなグラフとなる。

実験においては、B-GBI 法でチャンクするペアの選択には頻度、B-GBI 法が出力したペアの中で最も分岐に有効なペアを選択するには情報利得 [Quinlan 86] を利用した。GBI 法で抽出したパターンを属性として用いることの効果を調べるために、まず B-GBI 法のビーム幅を 1 として決定木は次の 2 つの方法で構築した：1) 根ノードでは n_r（1〜10）回チャンクし、その他のノードでは 1 回しかチャンクしない、2) 決定木の各ノードで n_e（1〜7）回チャンクする。図 3 に示す DT-GBI 法の終了条件を D 内のグラフ数が 10 以下とした。DT-GBI 法に沿って構築された決定木の予測誤り率を、両実験ともに 10-fold cross-validation で評価した。評価実験の結果を示す図 7 から分かるように、最も低い誤り率は：実験 1) で $n_r = 5$ とした場合の予誤り率は 9.43%、実験 2) で $n_e = 3$ とした場合の予測誤り率は 8.49%になった。また、書き換えられたグラフに残ったペアしか選ばない場合、実験 1) で n_r が 5 を、実験 2)n_e が 3 を超えるとチャンキングを繰り返すほど誤り率が悪くなってしまった。ペアが大きくなり過ぎると分類能力（情報利得など）が低下してしまうと考えられる。

さらに次の 2 つの方法で決定木を構築した：3)n_r を 5 と固定して探索ビーム幅を 2〜15 に変更した、4)n_e を 3 回と固定して探索ビーム幅を 2〜12 に変更した。実験 1) および 2) と同

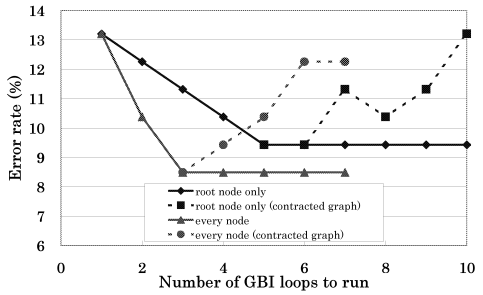

図 7: 実験 1) および 2) の結果

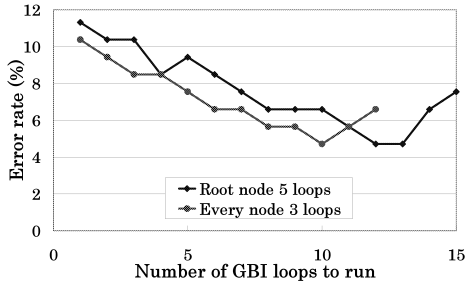

図 8: 実験 3) および 4) の結果

様に構築された決定木の予測誤り率を 10-fold cross-validation で評価した。その結果を図 8 に示す。実験 3) において、最も低い誤り率はビーム幅が 12 に達した時の 4.72%であり、実験 4) においても、最も低い誤り率はビーム幅が 10 に達した時の 4.72%であった。いずれの実験でも探索能力が改善されたことが確認できた。さらに同図より、ビーム幅が一定値より広くなると誤り率が上がってしまうことがあることが分かる。各ステップでのビーム幅が有限であり、かつ、ビーム幅を増やすと途中のステップで上位だった部分グラフが下位に下がってしまって選ばれなかったことが考えられる。

最後に、過学習を防ぐために枝刈りを導入することによる分類精度向上に対する実験を行った。本稿で枝刈りの対象としたは：5) 実験 3) において n_r を 5 およびビーム幅を 12 とした時の決定木、6) 実験 4) において n_e を 3 およびビーム幅を 10 とした時の決定木である。枝刈り方法は C4.5[Quinlan 93] と同様の pessimistic pruning を採用した。DT-GBI 法および枝刈りで決定木を構築して 10-fold cross-validation で予測誤り率を求めた結果、実験 5) の決定木の誤り率は 4.72%と変動しなかったが、実験 6) の決定木の誤り率は 3.77%にまで減少した。両実験の枝刈り後の決定木を図 9 および図 10 に示す。

5. おわりに

本稿では、グラフ構造データに対して属性と属性値を生成しながら決定木を構築する手法（DT-GBI 法）の探索能力をビーム探索の導入により改善した。実際のデータに使用し、ビーム幅をある程度まで広くすることで構築された決定木の予測精度が向上したことを確認した。

現段階では、それぞれの決定木のノードにおいて部分パターンを抽出するのに何回 B-GBI 法を実行させるかは手動で決めている。今後は分岐ペアに対する評価関数（情報利得）の変動を見ながら自動的に B-GBI 法の回数を決める手法を導入する予定である。

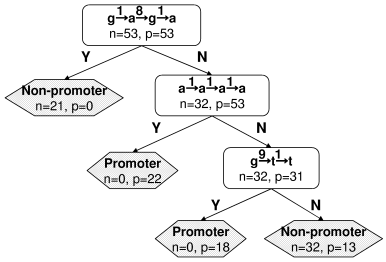

図 9: 実験 5) によって得られた決定木

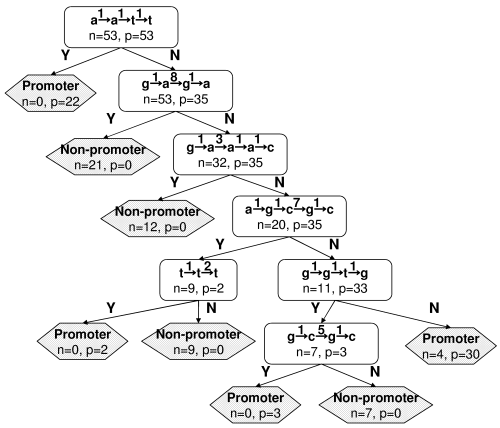

図 10: 実験 6) によって得られた決定木

参考文献

[Blake 98] C. L. Blake, E. Keogh, and C. Merz: UCI Repository of Machine Learning Database (1998), http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[Breiman 84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone: *Classification and Regression Trees*, Wadsworth & Brooks/ Cole Advanced Books & Software, 1984.

[Matsuda 02] T. Matsuda, H. Motoda, T. Yoshida, and T. Washio: Knowledge Discovery from Structured Data by Beam-wise Graph-Based Induction, in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence, Springler Verlag, LNAI2417*, pp. 255-264, 2002.

[Quinlan 86] J. R. Quinlan: Induction of decision trees, *Machine Learning*, Vol.1, pp.81-106, 1986.

[Quinlan 93] J. R. Quinlan: *C4.5: Programs For Machine Learning*, Morgan Kaufmann Publishers, 1993.

[Towell 93] G. G. Towell and J. W. Shavlik: Extracting refined rules from knowledge-based neural networks, *Machine Learning*, Vol.13, pp.71-101, 1993.

[ジアムサクン 02] ジアムサクン、松田、元田、鷲尾、吉田：Graph-Based Induction を用いたグラフ構造データに対する分類器の構築、2002 年度人工知能学会全国大会（第 16 回）論文集、セッション 1A3-02、2002。

[松田 01] 松田、元田、鷲尾：一般グラフ構造データにたいする Graph-Based Induction とその応用、人工知能学会誌、Vol.16、No.4、pp.363-374、2001。

[松田 02] 松田、元田、吉田、鷲尾：Graph-Based Induction による分類学習のための構造データからの属性構築、2002 年度人工知能学会全国大会（第 16 回）論文集、セッション 1A4-03 、2002。

[吉田 97] 吉田、元田：逐次ペア拡張に基づく昨日推論、人工知能学会誌、Vol.12、No.1、pp.58-97、1997。