1D3-3

遺伝的プログラミングと編集距離を利用した特徴的な VLDC 木パターンの獲得

Acquisition of Characteristic Tree Patterns with VLDC's by Genetic Programming and Edit Distance

中居 翔平[*1]　宮原 哲浩[*1]　久保山 哲二[*2]　内田 智之[*1]　鈴木 祐介[*1]

Shohei Nakai　Tetsuhiro Miyahara　Tetsuji Kuboyama　Tomoyuki Uchida　Yusuke Suzuki

[*1] 広島市立大学情報科学研究科

Graduate School of Information Sciences, Hiroshima City University

[*2] 学習院大学計算機センター

Computer Centre, Gakushuin University

Knowledge discovery from structured data is an important task in machine learning and data mining. We propose a learning method for acquiring characteristic tree patterns with VLDC's from positive and negative tree structured data by using Genetic Programming and tree edit distance. We report experimental results on applying our method to glycan data.

1. はじめに

木構造データからの機械学習やデータマイニングの研究が注目されている. 糖鎖は木構造をしており, 核酸(DNA)とタンパク質に続く 3 番目に重要な生体分子である. 遺伝的プログラミング(Genetic Programming, GP) [Poli 08]とは, 遺伝的アルゴリズム(GA)の遺伝子型を拡張し, 構造的表現(木構造)を扱えるようにしたものである.

遺伝的プログラミングと編集距離とを利用して正事例と負事例の木データから特徴的な VLDC 木パターン[Zhang 94]を獲得する提案手法[Nakai 13]に, 新たな糖鎖データを適用してその有効性を確認したので, 本稿で報告する. VLDC(variable-length don't care)は木データの一部を代入できる構造的変数である. 関連研究として, 遺伝的プログラミングによる特徴的なタグ木パターン獲得手法[Nagamine 07], 有向グラフ構造に対する進化的計算手法 [Katagiri 00]や TTSP グラフパターンの進化的獲得手法[Nagai 12]などがある.

2. 準備

本稿では, 木構造は順序木構造を持つものとし, 木構造データの構造的特徴を表現するため, VLDC 木パターンと呼ぶ木構造パターンを用いる. VLDC 木パターンは, ノードラベルでデータを表現し, 木データの一部を代入できる VLDC 変数を持つ. この VLDC 変数には Path-VLDC と Umbrella-VLDC の 2 種類がある. 木データの根から葉までの経路の一部が代入可能な VLDC 変数を Path-VLDC という. 表記では"|"で表される. 木データの根から葉までの経路の一部と, その経路の一部上のノードから出ているすべての部分木も代入可能な VLDC 変数を Umbrella-VLDC という. ただし経路で一番下のノードから出ている部分木は含まなくてもよい. 表記では"∧"で表される.

木データ T_1 と木データ T_2 の編集距離 treedist(T_1,T_2)は, T_1 を T_2 に変換するためにノード削除, ノード挿入, ノードラベル置換の 3 種類の編集操作を用いて編集する際にかかるコストの総和の最小値と定義する[Zhang 89]. S を VLDC 木パターン P における可能な VLDC 変数への代入の全ての集合とする. P の VLDC 変数に代入 s∈S を適用して得た木データを P(s)とする. P(s)と木データ T との編集距離が最小になるときの代入 s∈S による, P(s)と T の編集距離を, P と T の編集距離 treedist(P,T)と定義する[Zhang 94]. VLDC 木パターンと木データの編集距離の例を図 1 に示す. この例では, ノード削除, ノード挿入, ノードラベル置換の編集コストは全て 1 としている.

連絡先:宮原哲浩, 広島市立大学情報科学研究科, 〒731-3194
広島市安佐南区大塚東 3-4-1, miyares13@info.hiroshima-cu.ac.jp

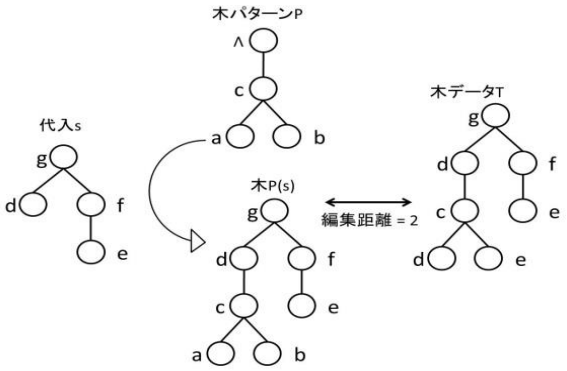

図 1: VLDC 木パターン P と木データ T の編集距離

3. GP と編集距離を利用した特徴的な VLDC 木パターンの獲得

以下で定義する, 特徴的 VLDC 木パターン獲得問題を対象とし, GP による獲得手法を述べる.

特徴的 VLDC 木パターン獲得問題:
入力:正事例と負事例からなる木データの有限集合 D
問題:D の特徴を表す VLDC 木パターンを獲得する.

GP-0 と GP-AUC という2つの, 木構造に基づく通常の GP による獲得手法を提案する. 異なる点は VLDC 木パターン P と木データ T がマッチすることの定義と, 個体である VLDC 木パターン P の適合度の定義であり, そのほかの設定は同じである. P が D の正事例集合にマッチする割合を P の正事例支持度という. P が D の負事例集合にマッチしない割合を P の負事例支持度という. (P の正事例支持度 + P の負事例支持度)/2 を P の総支持度という.

GP-0 では treedist(P,T)=0 である時だけ, P と T がマッチするという. P の総支持度を P の適合度と定義する. GP-AUC では編集距離の閾値 d に対して, treedist(P,T)≦d である時だけ, P と T がマッチするという. D の正事例集合に対する P にマッチする正事例の割合を P の真陽性率という. D の負事例集合に対する P にマッチする負事例の割合を P の偽陽性率という. d の値をずらすことで得られる P の真陽性率と偽陽性率から計算できる ROC 分析の AUC 値を P の適合度と定義する.

GP による特徴的 VLDC 木パターンを獲得する手法:
1. 正事例木データ集合から使用されているノードラベル, 親ノードと子ノードのラベルの関係, 木データのサイズ(ノードの個数)の最大値, 子の数の最大値をそれぞれ求める.

2. 1 で求めた値を基にランダムに初期 VLDC 木パターン集合を生成する.
3. VLDC 木パターンの適合度を求める.
4. 適合度の大きさに比例した確率によって VLDC 木パターンの選択を行う.
5. 交叉, 突然変異(部分木交換, 部分木追加, 部分木削除), 逆位, 複製の遺伝的操作により, 次世代の集団を生成する. (図 2 に VLDC 木パターンの交叉の例を示す.)
6. 終了条件である世代数まで達していれば終了する. そうでなければ 5 で生成された次世代の集団を現世代の集団として 3 へ戻る.

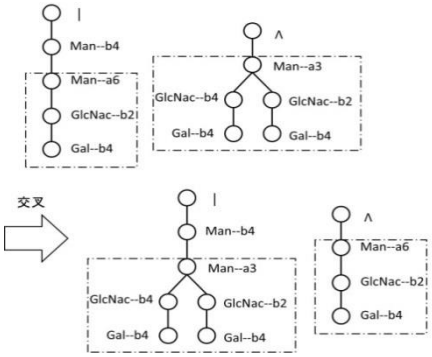

図 2:VLDC 木パターンの交叉例

4. 実験

3.で説明した 2 つの手法を用いて GP により適合度の高い VLDC 木パターンを獲得する実験を行った. 実験データとして KEGG GLYCAN データベースに登録されている次の 3 種類の糖鎖データを使用した. 白血病に関する正事例データ 177 個, 負事例データ 302 個. 結腸癌に関する正事例データ 87 個, 負事例データ 47 個. 嚢胞性線維症(CF 症)に関する正事例データ 89 個, 負事例データ 71 個. GP のパラメータは次の通りである. 個体数:50, 交叉確率:0.7, 突然変異確率:0.1, 逆位確率:0.1, 複製確率:0.1, 最大世代数:200, ルーレットサイズ:4, トーナメントサイズ:4, エリートサイズ:5. 編集コストは次のように設定した. ノード削除とノード挿入はどちらも 1 とする. ノードラベル置換の編集コストは, 糖ラベルが同じかつ結合ラベルも同じ時は 0, 糖ラベルだけまたは結合ラベルだけが同じ時は 0.5, 糖ラベルが異なりかつ結合ラベルも異なる時は 1 とする.

3 種類のデータを使用した時の 2 つの手法の最終世代の最良個体の適合度などの値(10 試行の平均値)を表 1 に示す. GP-0 では, 正事例支持度, 負事例支持度, 適合度, 総支持度の値は最終世代の最良個体の値である. GP-AUC では, 適合度の値は最終世代の最良個体の値である. さらに, 最終世代の最良個体の総支持度が最大となるように編集距離の閾値 d を定めたときの総支持度, 正事例支持度, 負事例支持度を示す. GP-0 と GP-AUC の適合度の推移(10 試行の平均値)を図 3 に示す. 図 4 に白血病に関するデータを与えたときの GP-AUC による最終世代の最良個体である VLDC 木パターンの例を示す.

5. おわりに

遺伝的プログラミングと編集距離を利用した特徴的な VLDC 木パターンを獲得する手法を, 3 種類の糖鎖データに適用してその有効性を確認した. 糖鎖データ以外の木構造データに提案手法を適用して有効性を検証する実験を行うことが課題として考えられる.

表 1:GP-0 と GP-AUC における最終世代の最良個体の比較

	GP-0			GP-AUC		
使用データ	白血病	結腸癌	CF 症	白血病	結腸癌	CF 症
正事例支持度	0.72	0.37	0.45	0.81	0.97	0.88
負事例支持度	0.92	0.95	0.92	0.90	0.77	0.78
適合度	0.82	0.66	0.68	0.91	0.93	0.86
総支持度	0.82	0.66	0.68	0.86	0.87	0.83

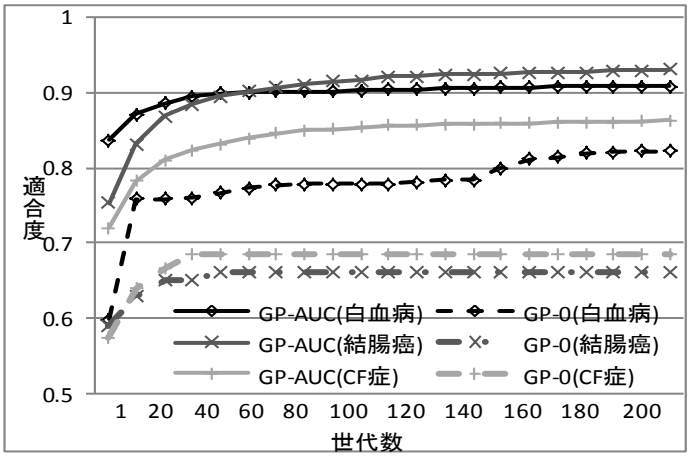

図 3: 各世代の適合度の平均値の推移

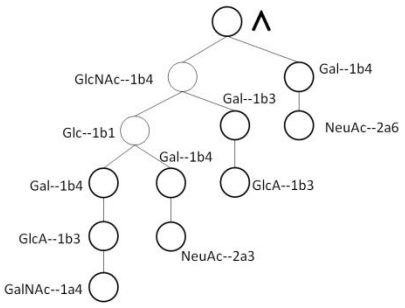

図 4: GP-AUC による最良個体である VLDC 木パターン

参考文献

[Katagiri 00] H.Katagiri et al., Genetic Network Programming - Application to Intelligent Agents, Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, pp.3829-3834, 2000.
[Nagai 12] S.Nagai et al., Acquisition of Characteristic TTSP Graph Patterns by Genetic Programming, Proc. 2012 IIAI International Conference on Advanced Applied Informatics, pp.340-344, 2012.
[Nagamine 07] M. Nagamine et al., A Genetic Programming Approach to Extraction of Glycan Motifs Using Tree Structured Patterns, Proc. AI 2007, Lecture Notes in Artificial Intelligence, Springer-Verlag vol.4830, pp.150-159, 2007.
[Nakai13] S.Nakai et al., Acquisition of Characteristic Tree Patterns with VLDC's by Genetic Programming and Edit Distance, Proc. 2013 IIAI International Conference on Advanced Applied Informat**ics**, pp.147-151, 2013.
[Poli 08] R.Poli et al., A Field Guide to Genetic Programming, Lulu Press, 2008.
[Zhang 89] K.Zhang and D. Shasha, Simple Fast Algorithms for the Editing Distance between Trees and Related Problems, SIAM Journal on Computing, Vol.18, No.6, pp.1245-1262, 1989.
[Zhang 94] K.Zhang et al., Approximate Tree Matching in the Presence of Variable Length Don't Cares, Journal of Algorithms Vol.16, No.1, pp.33-66, 1994.