2I3-5

VLDC 木パターン集合を個体とする進化的手法による複合的木構造パターンの獲得

Acquisition of Multiple Tree Structured Patterns by an Evolutionary Method using Sets of Tree Patterns with VLDC's as Individuals

中居 翔平[*1]　宮原 哲浩[*1]　鈴木 祐介[*1]　久保山 哲二[*2]　内田 智之[*1]
Shohei Nakai　Tetsuhiro Miyahara　Yusuke Suzuki　Tetsuji Kuboyama　Tomoyuki Uchida

[*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 multiple tree structured patterns from positive and negative tree structured data by an evolutionary method using sets of tree patterns with VLDC's as Individuals. We report experimental results on applying our method to glycan data.

1. はじめに

遺伝的プログラミング(Genetic Programming, GP)[Koza 92][Poli 08]とは，遺伝的アルゴリズムの遺伝子型を拡張し，木構造のような構造的表現を扱えるようにした進化的手法である．木構造データの例として糖鎖データがある．糖鎖は核酸(DNA)とタンパク質に続く 3 番目に重要な生体分子である．その構造の複雑さから糖鎖の機能や構造の解析は核酸やタンパク質に比べて進んでいない．

本稿では，VLDC 木パターンと木データの編集距離[Zhang 94]と遺伝的プログラミングを利用して，正事例集合と負事例集合の木データから特徴的な VLDC 木パターン集合を獲得する進化的手法[Nakai 14c]を報告する．この手法は，正事例集合と負事例集合の木データから特徴的な単一 VLDC 木パターンを獲得する手法[Nakai 13][中居 14b]を利用するものであり，[中居 14a]の手法を発展させて VLDC 木パターン集合を個体とする進化的手法としたものである．VLDC(variable-length don't care)と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，木データの正事例集合からの木パターン和の多項式時間学習[Arimura 93]，有向グラフ構造に対する進化的計算[Katagiri 00]，木データの正事例集合と負事例集合からの遺伝的プログラミングによる特徴的タグ木パターンの獲得[Nagamine 07]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．以降 VLDC 木パターンを木パターンということもある．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できる 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できる 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"Λ"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して定義される[Zhang 89]．S を木パターン P における VLDC 変数への可能な代入すべての集合とする．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．木パターン P と木データ T の編集距離 treedist(P,T)を，P(s)と T との編集距離 treedist(P(s),T)が最小になるときの P(s)と T の編集距離，すなわち treedist(P,T)=min{treedist(P(s),T)|s ∈S}と定義する[Zhang 94]．木パターン P と木データ T の編集距離の例を図 1 に示す．ここでは，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すべて 1 としており，treedist(P,T)=2 となる．

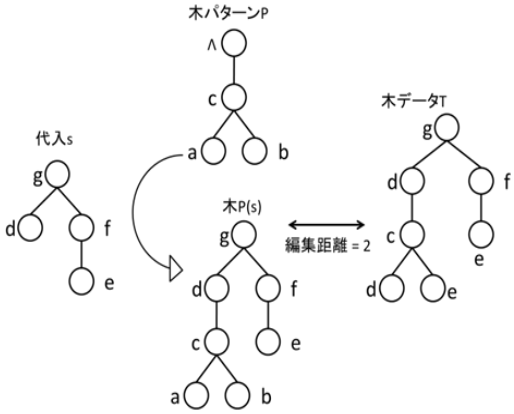

図 1:木パターン P と木データ T の編集距離

3. VLDC 木パターン集合を個体とする進化的手法による複合的木構造パターンの獲得

本研究では GP-0 と GP-AUC の 2 つの個体評価法を用いて特徴的 VLDC 木パターン集合の獲得を行う．提案する特徴的 VLDC 木パターン集合獲得手法は，特徴的単一 VLDC 木パターン獲得手法[Nakai 13][中居 14b]を基にしている．まず，単一 VLDC 木パターン獲得手法について説明する．

個体には VLDC 木パターンを用いる．2 つの個体評価法 GP-0 と GP-AUC における，木パターンが木データにマッチすることの定義と木パターンの適合度の定義を説明する．

GP-0 では VLDC 木パターン P と木データ T の編集距離が 0 の時に P と T がマッチするという．P の適合度は(D の正事例集合に P がマッチする割合 + D の負事例集合に P がマッチしない割合)/2 と定義する．

GP-AUC では VLDC 木パターン P と木データ T の編集距離が閾値 d 以下の時に P と T がマッチするという．適合度計算の

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，miyares14@info.hiroshima-cu.ac.jp

ために次の値を定義する. 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. P の適合度は 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 値と定義する. 単一 VLDC 木パターン獲得問題を以下に定義する.

入力:正事例集合と負事例集合からなる木データの有限集合 D
問題:D に関する適合度の高い VLDC 木パターンを獲得する.

特徴的な単一 VLDC 木パターン獲得手法を以下に示す.

特徴的な単一 VLDC 木パターン獲得手法

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, 木データのサイズの最大値, 子の数の最大値を求める.
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. 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.

次に VLDC 木パターン集合獲得手法について説明する. 個体には VLDC 木パターン集合を用いる. 2 つの個体評価法 GP-0 と GP-AUC における, 木パターン集合が木データにマッチすることの定義と木パターン集合の適合度の定義を説明する.

GP-0 では VLDC 木パターン集合Πに含まれる VLDC 木パターンの少なくとも 1 つと木データ T の編集距離が 0 の時にΠが T にマッチするという. Πの適合度は(D の正事例集合にΠがマッチする割合 + D の負事例集合にΠがマッチしない割合)/2 と定義する.

GP-AUC では VLDC 木パターン集合Π = { $\pi_1, \pi_2, \ldots, \pi_n$} と木データ T の編集距離は, Πに含まれる VLDC 木パターンのうち T との編集距離が最も小さい VLDC 木パターンとの編集距離, すなわち min{treedist(π_i,T)｜1≦i≦n}と定義する. ΠとT

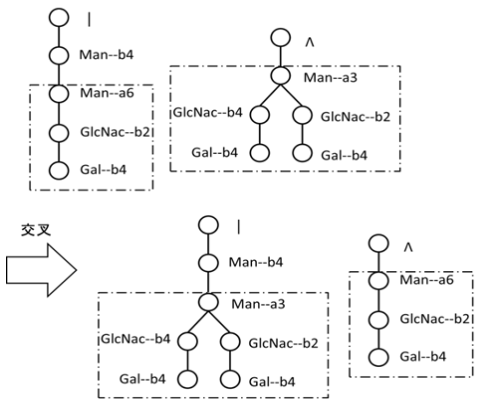

図 2:木パターンの交叉例

の編集距離が閾値 d 以下の時にΠと T がマッチするという. 適合度計算のために次の値を定義する. D の正事例集合に対するΠにマッチする正事例の割合をΠの真陽性率という. D の負事例集合に対するΠにマッチする負事例の割合をΠの偽陽性率という. VLDC 木パターン集合Πの適合度は d の値をずらすことで得られるΠの真陽性率と偽陽性率から計算できる ROC 分析の AUC 値と定義する. VLDC 木パターン集合獲得問題を以下に定義する.

入力:正事例集合と負事例集合からなる木データの有限集合 D
問題:D に関する適合度の高い VLDC 木パターン集合を獲得する.

本稿で提案する, 特徴的な VLDC 木パターン集合獲得手法を以下に示す. まず, その概要と GP での適合度計算法を説明する. VLDC 木パターン集合を個体とする進化的手法(メインルーチン)では単一 VLDC 木パターンを個体とする GP をサブルーチンとして使う. メインルーチンの進化的手法では, 個体である VLDC 木パターン集合{ $\pi_1, \pi_2, \ldots, \pi_n$}を VLDC 木パターン列[$\pi_1, \pi_2, \ldots, \pi_n$]として表現する. s 世代の GP の終了時における単一 VLDC 木パターンπの適合度をπの基本適合度といい(ステップ 7), 進化的手法で獲得した適合度の高い VLDC 木パターン集合でのπの出現回数に比例した値をπの加算適合度という. 加算適合度の高い VLDC 木パターンπは, 高い適合度を持つ VLDC 木パターン集合の良い構成要素であるので, πの基本適合度にπの加算適合度を加えた値をπの適合度として(ステップ 10), s+1 世代の GP を開始する.

特徴的な VLDC 木パターン集合獲得手法

1. 正事例のクラスタ数(c), 単一 VLDC 木パターンの上位数(k), 進化的手法の集団サイズ(b), 進化的手法のエリートサイズ(e), GP の集団サイズ(b'), GP のエリートサイズ(e'), 最大世代数(n), VLDC 木パターンの加算適合度の最大値(C_{add}) を設定する.
2. 初期世代として s=1 とする.
3. Pos を D の正事例全体の集合とし, Neg を D の負事例全体の集合をとする.
4. 編集距離と Fuzzy c-means 法を用いて Pos のクラスタリングを行い, Pos を c 個の部分集合 $Pos_1,\ldots,Pos_c$ に分類をする (Pos=$Pos_1 \cup Pos_2 \cup Pos_3 \cup \ldots \cup Pos_c$). Pos_jを正事例集合, Neg を負事例集合とするデータの与え方をD_j (1≦j≦c)とする.
5. $D_1, D_2, \ldots, D_c$それぞれに対して, 初期 VLDC 木パターン集合を生成し, 特徴的な単一 VLDC 木パターンを獲得する GP による学習過程を始める. それぞれの学習過程を $GPL_1, GPL_2, \ldots, GPL_c$とする.
6. 現世代(s)の$GPL_1, GPL_2, \ldots, GPL_c$を実行する.
7. $GPL_1, GPL_2, \ldots, GPL_c$の個体の VLDC 木パターンの適合度を評価して, 基本適合度とする.
8. PAT_{seq}^{prv}は前の世代(s-1)の適合度上位 e 個の VLDC 木パターン列の集合とする. PAT_{seq_best}を$PAT_{seq} \cup PAT_{seq}^{prv}$の適合度上位 b 個の VLDC 木パターン列の集合とする. ここで, PAT_{seq}を現世代(s)の

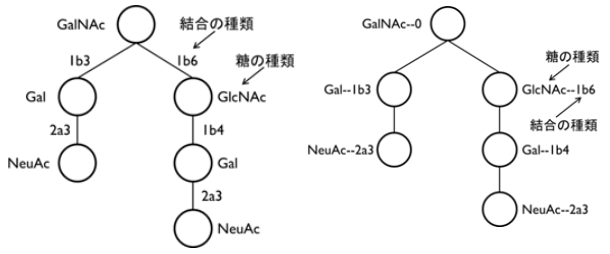

図 3: KEGG GLYCAN データベースでの表記方法の例(左図)と本研究での表記方法の例(右図)

各GPL_jの適合度が上位 k 個の VLDC 木パターンπ_j (1≤j≤c) から成る VLDC 木パターン列$[\pi_1, \pi_2, \ldots, \pi_c]$のすべてから成る集合とする.

9. 終了世代(s=n)に達していれば終了とする.
10. GP の学習過程GPL_jの各 VLDC 木パターンπ_jに対し, PAT_{seq_best}の VLDC 木パターン列の j 番目の要素としての, VLDC 木パターンπ_jの出現回数を n(π_j)とする. GPL_jの中の VLDC 木パターンπ_jの基本適合度に π_jの加算適合度($C_{add} * \frac{n(\pi_j)}{b}$)を加えた値を$\pi_j$の適合度とする.
11. GP の処理過程$GPL_1, GPL_2, \ldots, GPL_c$を継続し, 現世代(s)の集団を次世代(s+1)の集団とし, s=s+1 として 6 へ戻る.

4. 実験

3 節で説明した VLDC 木パターン集合を個体とする進化的手法を実装し, 実データを用いた特徴的な VLDC 木パターン集合を獲得する評価実験を行った.

実験データとして KEGG GLYCAN データベース[Hashimoto 03]の結腸癌に関する正事例 87 個, 負事例 47 個の糖鎖データを使用した. 実験で扱う糖鎖データの表記方法と編集距離の計算に関係する本研究でのノードラベル置換コストの定義について説明する. 図 3 に KEGG GLYCAN データベースでの表記方法の例と本研究での表記方法の例を示す. 本研究では一つのノードで糖の種類と結合の種類の両方を表現するように変換する. 結合の種類はエッジを構成するノードのうち葉に近いノードに追加する. 結合の種類に 0 があるのは変換前にエッジがなかったことを示す.

ノードラベル置換コストは次のように設定した. (a)糖ラベルが同じかつ結合ラベルも同じ時, ノードラベル置換コスト:0.0. (b)糖ラベルだけまたは結合ラベルだけが同じ時, ノードラベル置換コスト:0.5. (c)糖ラベルが異なりかつ結合ラベルも異なる時, ノードラベル置換コスト:1.0.

特徴的 VLDC 木パターン集合を獲得する進化的手法(メインルーチン)のパラメータは次のように設定した. 集団サイズ(b):50, エリートサイズ(e):5, 最大世代数(n):200, 正事例のクラスタ数(c):3, 単一 VLDC 木パターンの上位数(k):5, VLDC 木パターンの加算適合度の最大値(C_{add}):0.1 または 0.2.

GP による単一 VLDC 木パターン獲得手法(サブルーチン)のパラメータは次のように設定した. 集団サイズ(b'):50, エリートサイズ(e'):5, 最大世代数(n):200, 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, トーナメントサイズ:4.

特徴的な VLDC 木パターン集合獲得手法において, 個体評価法 GP-0,GP-AUC を用いる設定を, それぞれ GP-0-set, GP-AUC-set とする. 加算適合度の最大値(C_{add})も示す. 比較実験として行った. 特徴的な単一 VLDC 木パターン獲得手法において, 個体評価法 GP-0,GP-AUC を用いる設定を, それぞれ GP-0-single, GP-AUC-single とする. それぞれの設定において, 適合度の高い VLDC 木パターン集合または単一 VLDC 木パターンを獲得する実験を 10 試行ずつ行った. 2 つの個体評価法で適合度の定義が異なるので, 得られた個体を比較するため (正事例集合にマッチする割合+負事例集合にマッチしない割合)/2 で定義される総支持度を定義し, 個体を比較することにする. GP-0 における個体の適合度は総支持度のことである.

表 1: 最終世代の最良個体の比較

	GP-0-single	GP-0-set (Cadd=0.1)	GP-0-set (Cadd=0.2)
適合度	0.661	0.785	0.800
総支持度	0.661	0.785	0.800
ノード数	2.8	8.3	9.1
	GP-AUC-single	GP-AUC-set (Cadd=0.1)	GP-AUC-set (Cadd=0.2)
適合度	0.935	0.923	0.925
総支持度	0.888	0.880	0.870
距離の閾値	14.6	8.3	6.9
ノード数	21.9	49.7	44.3

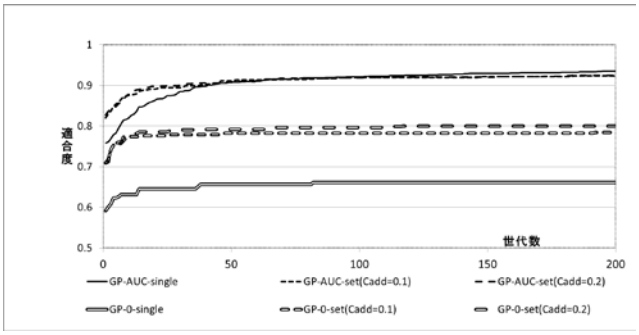

図 4: 各世代の適合度の平均値の推移

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-single,GP-0-set では, 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-single, GP-AUC-set 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 d の値(距離の閾値)と総支持度を示す. 適合度の推移(10 試行の平均値)を図 4 に示す. 図 5 に GP-0-set(C_{add}=0.2)による 10 試行の最終世代の最良個体である VLDC 木パターン集合を示す. 図 6 に GP-AUC-set(C_{add}=0.2)による 10 試行の最終世代の最良個体である VLDC 木パターン集合と距離の閾値を示す.

表 1 と図 4 より, GP-0-single, GP-0-set(C_{add}=0.1), GP-0-set(C_{add}=0.2)の順に, 適合度が高くなっていくことがわかる. GP-AUC-single, GP-AUC-set(C_{add}=0.1), GP-AUC-set(C_{add}=0.2)では, 適合度と総支持度はほぼ等しいが, この順に距離の閾値が小さくなっていくことがわかる.

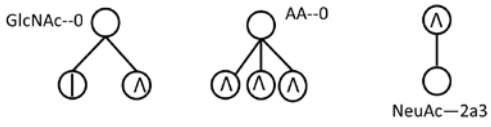

図 5: GP-0-set(C_{add}=0.2)の最良個体

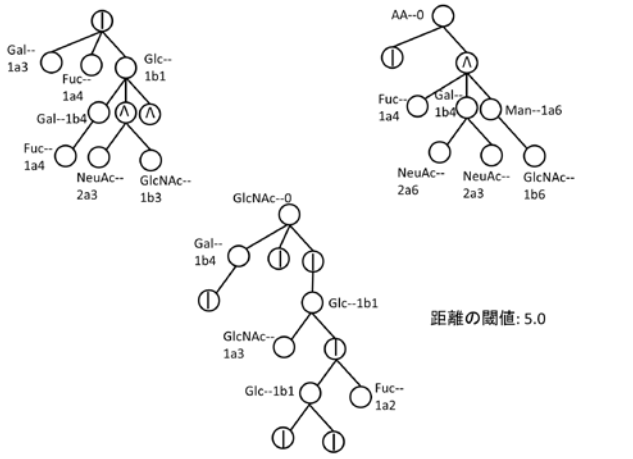

図 6: GP-AUC-set (C_{add}=0.2)の最良個体

5. おわりに

VLDC 木パターン集合を個体とする進化的手法による複合的木構造パターンを獲得する手法を提案した. 結腸癌に関する糖鎖データに適用してその有効性を確認した. 今後の課題として, 他の木構造データの適用, データごとに適切なパラメータを設定することなどが挙げられる.

参考文献

[Arimura 93] H. Arimura et al., Polynomial Time Algorithm for Finding Finite Unions of Tree Pattern Languages, Proc. NIL-91, Springer-Verlag, LNAI 659, pp.118–131, 1993.

[Hashimoto 03] K. Hashimoto et al., GLYCAN: The Database of Carbohydrate Structures, Genome Informatics, Vol.14, pp.649-650, 2003.

[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.

[Koza 92] J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, 1992.

[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.

[Nakai 13] 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 Informatics, pp. 113 - 118, 2013.

[中居 14a] 中居翔平ほか, 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 木パターン集合の獲得, 火の国情報シンポジウム 2014 論文集,3B-3, 2014.

[中居 14b] 中居翔平ほか, 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, 第 28 回人工知能学会全国大会論文集, 1D3-3, 2014.

[Nakai 14c] S.Nakai et al., Acquisition of Characteristic Sets of Tree Patterns with VLDC's Using Genetic Programming and Edit Distance, Proc. 2014 IEEE 7th International Workshop on Computational Intelligence and Applications, pp.147-151, 2014.

[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.