様式 C-19

科学研究費補助金研究成果報告書

平成　22 年 6 月 18 日現在

研究種目：基盤研究（C）
研究期間： 2007～2009
課題番号：１９５００１２９
研究課題名（和文）　データマイニングと機械学習による半構造データからの情報融合

研究課題名（英文）Information Fusion from Semi-structured Data using Data Mining and Machine Learning

研究代表者
宮原　哲浩（MIYAHARA TETSUHIRO）
広島市立大学・情報科学研究科・准教授
研究者番号：９０２０９９３２

研究成果の概要（和文）：知識発見と情報融合を実現するため，半構造データからのデータマイニングと機械学習について研究した．厳密には定義されていない構造を持つデータを半構造データという．主に，半構造データとして木構造で表される糖鎖データを対象とし，その構造的特徴を表す木構造パターンを獲得する機械学習手法を提案した．手法として，木構造などの構造的表現を扱うことのできる進化的最適解探索手法である遺伝的プログラミングを用いた．

研究成果の概要（英文）：We have studied data mining and machine learning from semi-structured data for knowledge discovery and information fusion. As main results, we have proposed machine learning methods for acquiring characteristic tree structured patterns from glycan data as semi-structured data. The methods are based on genetic programming for evolving solutions from structured data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,300,000	390,000	1,690,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：データマイニング，機械学習，木構造データ，遺伝的プログラミング

１．研究開始当初の背景

コンピュータ上で作成されたコンテンツやネットワーク上で収集・蓄積されたデータが爆発的な勢いで増加している．流通するコンテンツや収集されるデータの規模が，人間がそれに隠されている知識を引き出す能力をはるかに超えるようになっている．大規模で多様なコンテンツやデータから，真に必要な情報を取り出して，取り出した情報を統合・融合するための手法やシステムの研究開発が求められるようになっている．

２．研究の目的

本研究課題で対象とするのは，半構造データ(semi-structured data)であり，Web 文書，Web データ，HTML/XML ファイルなどがそ

の代表である．このようなデータは，文字列データよりは構造化されているが，関係データベースのような厳密な表形式の構造を持たないという意味で，半構造データと呼ばれている．本研究では，データマイニングと機械学習技術を活用して，大規模な半構造データからの情報抽出と情報融合を実現するための技術を開発することを目的とする．更には，分子生物データなど様々な分野における構造を持つデータからの知識発見，情報融合を目指す．

３．研究の方法

(1)糖鎖は木構造をしており，本研究課題の対象として適していることがわかった．そこで本研究では，半構造データとして木構造で表される糖鎖データを対象とし，その構造的特徴を表す木構造パターンを獲得する機械学習手法を開発した．以下では主にこの研究について説明する．

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，各種の糖が結合した一群の化合物であると言われている．糖鎖は木構造をしており，その構造は多様である．そこで，本研究では木構造パターンを用いた糖鎖データからの構造的特徴抽出手法を提案する．探索には，木構造などの構造的表現を扱うことのできる進化的最適解探索手法である遺伝的プログラミング(Genetic Programming, GP)を用いる．

本研究では，木構造データのパターンを表現するためにタグ木パターンを用いる．タグ木パターンとは，任意の木を代入できる構造的変数，任意の辺ラベルを表現できるワイルドカード(?)，辺ラベルとしてキーワードを持つ木構造パターンである．

タグ木パターン π が木構造データ T にマッチするとは， π の変数に木を代入し，ワイルドカードを辺ラベルで置き換えた後の木が T と一致するときにいう．図１に，糖鎖データに対応する木構造データ t_1，t_2 とタグ木パターン π の例を示す．π は t_1 にはマッチするが，t_2 にはマッチしない．

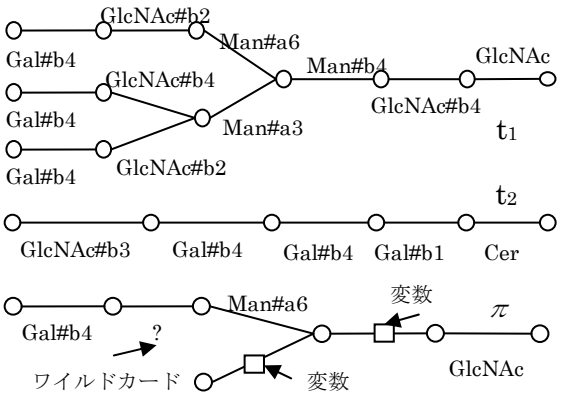

図１：糖鎖データとタグ木パターン

本研究では単一タグ木パターン発見問題とタグ木パターン集合発見問題の２つの問題を対象とした．

(2)単一タグ木パターン発見問題：
入力：正事例と負事例からなる木構造データの有限集合 D
問題：D の多くの正事例にマッチし，負事例にあまりマッチしない特徴的なタグ木パターン π を発見する．

単一タグ木パターン発見問題に対する GP に基づく手法は次のとおりである．
(i)D の正事例から，タグ木パターンで使用するキーワードの有限集合 KW を求める．
(ii)KW に含まれるキーワードを辺ラベルとして用いてランダムに初期タグ木パターン集合を発生させる．
(iii)タグ木パターンの適合度を求める．
(iv)適合度の大きさに比例した確率によってタグ木パターンの選択を行う．
(v)遺伝的操作により，次世代の集団を生成する．
(vi)終了条件が満たされているときは終了．そうでなければ(iii)へ戻る．

タグ木パターンの支持度（データを正しく説明する割合）と具体化度（どのくらい具体的であるかを示す数値）から適合度を定める．タグ木パターンに対する遺伝的操作として，交叉，部分木変更，部分木追加，部分木削除，辺ラベル変更を用いる．最初の 4 つの操作は，タグ木パターンの変数を特別な辺とみなした時の，木に基づく GP の遺伝的操作である．交叉の適用例を図２に示す．

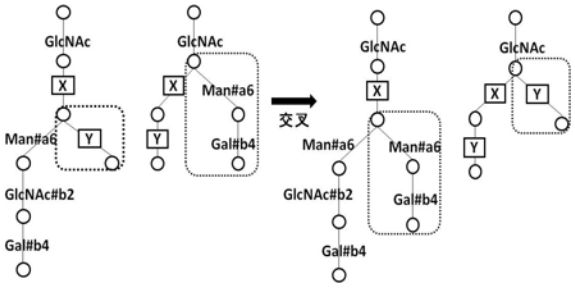

図２：タグ木パターンに対する交叉

(3)タグ木パターン集合発見問題：
入力：正事例と負事例からなる木構造データの有限集合 D
問題：D の多くの正事例にマッチし，負事例にあまりマッチしない特徴的なタグ木パターン集合 Π を発見する．

ここで，タグ木パターンの集合 Π が木 T にマッチするとは，Π に含まれるタグ木パターンのいずれかが，T にマッチすることをいう．

タグ木パターン集合発見問題に対する GP に基づく手法は次のとおりである．

(i)木の編集距離と k-means 法により，D の正事例集合をクラスタリングし，部分集合 $p_1,\ldots,p_k$ に分割する．
(ii)D の負事例集合と p_i の和集合を d_i とする．d_i からキーワード集合 kw_i を決定し，k 個の GP を部分過程として並列的に実行する．
(iii) 終了条件が満たされた時に d_i に対して適用された i 番目の GP 部分過程における適合度が最大の個体を π_i とし，$\Pi=\{\pi_1,\ldots,\pi_k\}$ を出力する．

タグ木パターン π がマッチする木全体の集合を $L(\pi)$ で示す．これは π の説明能力を表す．タグ木パターン集合 $\{\pi_1,\ldots,\pi_k\}$ の説明能力は $L(\{\pi_1,\ldots,\pi_k\}) = L(\pi_1) \cup\cdots\cup L(\pi_k)$ と定義する．図３はタグ木パターン集合発見問題を解く手法における正事例のクラスタリングの効果を表す．

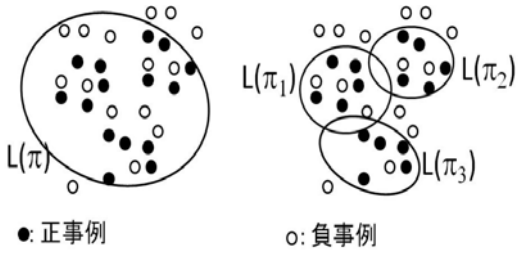

図３：正事例のクラスタリングとタグ木パターン集合の説明能力

４．研究成果

(1)4 つの主要な血液成分である，白血球，赤血球，血漿，血清に関する糖鎖データを対象として実験を行った．

(2)単一タグ木パターン発見問題では，白血球に関与する糖鎖データ 176 個を正事例とし，それ以外のデータ 304 個を負事例とする．このデータから特徴的な単一のタグ木パターンを生成する試行を 10 回行った．各試行各世代において，適合度の値が最も大きい個体を最良個体とし，最終世代の最良個体を得た試行を最良試行とする．図 4 は最良個体の適合度，支持度，具体化度の，最良試行(実線で示す)と，10 試行の平均(点線で示す)の各世代における推移を示す．

図 4 より，全ての値が収束していることから，支持度と具体化度を用いて適合度を設定しても，正常に探索を終了することができていることがわかる．また，図 5 に示すタグ木パターンは最良試行の最良個体である．このタグ木パターンから，GlcNAc が根となり，これに結合するいくつかの単糖があり，さらに葉の付近に GlcNAc と Gal が結合している構造が，白血球に関与する糖鎖の構造には多いことがわかる．また，この最良個体の構造は先行研究で発見された白血球に関与する糖鎖データに頻出する部分木を含むようなものであった．

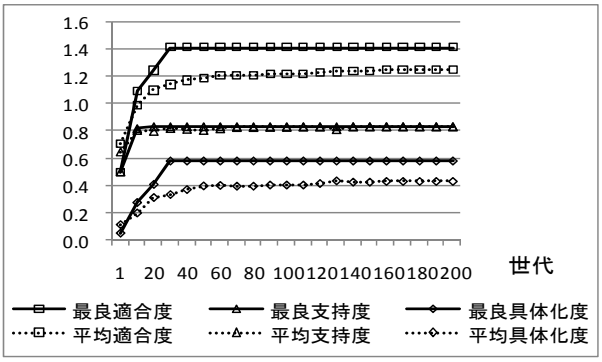

図 4：生成したタグ木パターンの適合度

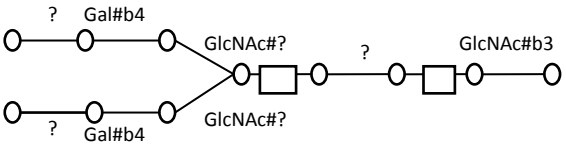

図 5.最良試行の最良個体の構造

(3)タグ木パターン集合発見問題では，赤血球に関与する糖鎖データ 164 個を正事例とし，それ以外のデータ 316 個を負事例とする．実験は，正事例集合の分類数 k を 1,3,4,5,10 と変化させて，それぞれ 3 試行ずつ行った．タグ木パターン集合がデータを正しく説明する割合を統合支持度という．

図 6 の統合支持度 k のグラフは，分類数が k の時の最良試行の統合支持度の推移を示す．これより，分類した時の方が値が大きくなっていることがわかる．タグ木パターン発見問題において，最も重要な尺度である支持度の値が大きくなったことから，正事例を分割して GP を行うことの効果を確認した．

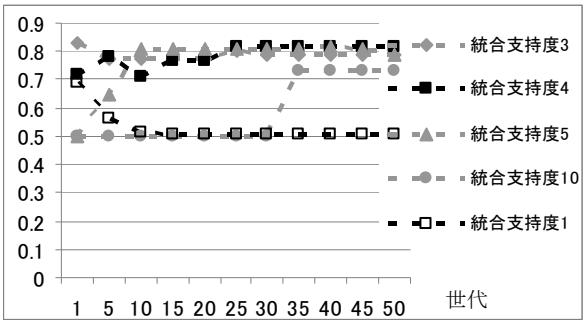

図 6：最良試行における統合支持度と分割数

(4)構造的変数を有するグラフパターンの集合に対する多項式時間質問学習可能性について研究を行った．具体的には，木あるいは TTSP グラフでモデル化できる半構造データ S に対して，多項式回の問合せで，S を生成可能なグラフパターンの集合を同定する多項式時間学習アルゴリズムを提案した．

５．主な発表論文等
（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 12 件）

① K.Yoshida, T.Miyahara, T.Kuboyama, Evolution of Multiple Tree Structured Patterns using Soft Clustering, Proceedings of 2nd International Conferencae on Computer and Automation Enginering, vol.5, 2010, pp.749-753.（査読有）

② M.Nagamine, T.Miyahara, T.Kuboyama, H.Ueda, K.Takahashi, Evolution of Multiple Tree Structured Patterns from Tree-Structured Data using Clustering, Proc. AI 2008, Lecture Notes in Artificial Intelligence, Springer-Verlag, vol.5360, 2008, pp.500-511.（査読有）

③ M.Nagamine, T.Miyahara, T.Kuboyama, H.Ueda, K.Takahashi, 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, 2007, pp.150-159.（査読有）

④ R.Okada, S.Matsumoto, T.Uchida, Y.Suzuki, T.Shoudai, Exact Learning of Finite Unions of Graph Patterns from Queries, Proc. ALT 2007, Lecture Notes in Artificial Intelligence, Springer-Verlag, vol.4754, 2007, pp.298-312.（査読有）

〔学会発表〕（計 3 件）

① 吉田 健吾，宮原 哲浩，久保山 哲二，ソフトクラスタリングを用いた複合的木構造パターンの進化的獲得，人工知能学会・人工知能基本問題研究会，2010 年 3 月 18 日，北海道大学

② 長嶺 将俊，宮原 哲浩，久保山 哲二，上田祐彰， 高橋健一，クラスタリングを用いた複合的木構造パターンの進化的獲得，人工知能学会・人工知能基本問題研究会，2009 年 3 月 14 日，学習院大学

６．研究組織

(1)研究代表者

宮原　哲浩（MIYAHARA TETSUHIRO）
広島市立大学・情報科学研究科・准教授
研究者番号：９０２０９９３２

(2)研究分担者

内田　智之（UCHIDA TOMOYUKI）
広島市立大学・情報科学研究科・准教授
研究者番号：７０２６４９３４

久保山　哲二（KUBOYAMA TETSUJI）
学習院大学・計算機センター・准教授
研究者番号：８０３０２６６０

廣渡　栄寿（HIROWATARI EIJU）
北九州市立大学・基盤教育センター・教授
研究者番号：６０２７４４２９
(H20～H21：連携研究者)

(3)連携研究者

該当なし