様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25 年 3 月 31 日現在

機関番号：２５４０３
研究種目：基盤研究（C）
研究期間：　2010～2012
課題番号：２２５００１３５
研究課題名（和文）構造的データに潜む知識を効果的に発見するためのデータマイニングと機械学習
研究課題名（英文）Effective Discovery of Hidden Structured Knowledge 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）：構造的データに潜む知識を効果的に発見するためのデータマイニングと機械学習について研究した．主な結果として，正事例と負事例から特徴的な TTSP グラフ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，構造的表現を扱うことのできる進化的探索手法である遺伝的プログラミングを用いた．構造的データから知識を発見するための他の学習手法も提案した．

研究成果の概要（英文）：We have studied effective discovery of hidden structured knowledge using data mining and machine learning. As a main result, we have proposed a machine learning method for acquiring characteristic TTSP graph patterns from positive and negative data. The method is based on genetic programming for evolving solutions from structured data. Also we have proposed other learning methods for discovering knowledge from structured data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,300,000	390,000	1,690,000
2011 年度	1,100,000	330,000	1,430,000
2012 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，グラフ構造データ，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

高速ネットワークで接続された世界中のコンピュータに膨大なデータや情報が蓄積されるようになっている．このようにして蓄積された膨大なデータから，ユーザや専門家が真に必要とする情報を取り出して，取り出した情報を知識として統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．近年では，構造化された多様なデータが，データマイニングの対象となってきており，伝統的なトランザクションデータベースや関係データベースからのデータマ

イニングだけでなく，グラフ構造データ，木構造データのような構造的データからのデータマイニングが大きな関心を集めるようになっている．

２．研究の目的

本研究課題では構造的データに潜む知識を効果的に発見するためのデータマイニングと機械学習について研究を行った．本研究の目的は，非均質で構造化された大規模なデータに潜む多様な知識の発見に焦点を当てて，必要とされる構造的知識を発見するためのデータマイニングと機械学習における新しい手法を開発することである．構造的データとしてグラフ構造データおよび木構造データを対象にした新たな機械学習手法を開発することを目的とする．

３．研究の方法

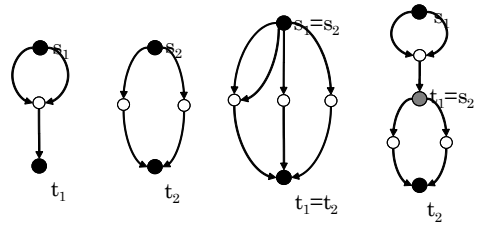
以下では,正事例と負事例から特徴的なTTSP グラフパターン(TTSP 項グラフという)を獲得する機械学習手法について述べる．

TTSP グラフと TTSP 項グラフについて説明する．TTSP グラフ（Two-Terminal Series Parallel graph）は，電気ネットワークやスケジューリングをコンピュータで扱う際にデータモデルとして用いられることが多い．次の(1),(2)で帰納的に定義される多重辺を許す有向グラフを TTSP グラフという．

(1)2 つの頂点u,vとuからvへの 1 つの辺から成る有向グラフは，ソースとしてuを持ちシンクとしてvを持つTTSPグラフである．(2)G_1, G_2をソースs_1，s_2とシンクt_1，t_2をそれぞれ持つTTSPグラフとする．このとき，s_1とs_2，t_1とt_2を同一視する操作を並列操作（得られるグラフは$G_1//G_2$と表現する），s_2とt_1を同一視する操作を直列操作（得られるグラフはG_1*G_2と表現する）という．並列操作，直列操作のどちらかによって得られるグラフはTTSPグラフである．この操作を図 1 に示す．

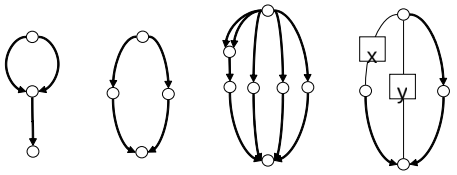
TTSPグラフのいくつかの辺を変数にしたものをTTSP項グラフという．この変数はTTSPグラフを代入できる構造的な変数である．TTSP項グラフgとTTSPグラフGに対して，gの変数を適切なTTSPグラフで置き換えることでG が得られるならば，gとGはマッチするという．例えば，図 2 において，TTSP項グラフgの変数xをTTSPグラフG_1で,変数yをTTSPグラフG_2で置き換えるとTTSPグラフGが得られるので，gとGはマッチする．

進化的手法による特徴的な TTSP 項グラフの獲得について説明する．構造的表現を扱うことのできる進化的探索手法である遺伝的プログラミング(Genetic Programming,以下では GP)を用いる．以下で定義する問題を対象とする．



G_1　　G_2　　$G_1//G_2$　　G_1*G_2

図 1　TTSPグラフG_1, G_2, $G_1//G_2$, G_1*G_2



G_1　　G_2　　G　　g

図 2　TTSPグラフG_1, G_2, G, TTSP項グラフg

TTSP 項グラフ発見問題：
入力：正事例と負事例からなる TTSP グラフの有限集合 D
問題：適合度の高い TTSP 項グラフ g を発見する

ここで，TTSP 項グラフ g の適合度を（g が D の正事例にマッチする割合 + g が D の負事例にマッチしない割合）/ 2 と定義する．適合度の高い TTSP 項グラフは，D の多くの正事例にマッチし，D の負事例にあまりマッチしないような，特徴的な TTSP 項グラフであると考えられる．

TTSP 項グラフ発見問題に対する GP の手順を次のようにする．GP の過程では TTSP 項グラフを個体として遺伝操作を行う．

(1)D の正事例から，TTSP 項グラフで使用するキーワードの有限集合 KW を求める．
(2)KW のキーワードを辺ラベルとしてランダムに初期 TTSP 項グラフを生成する．
(3)TTSP 項グラフの適合度を求める．
(4)適合度の大きさに比例した確率によって TTSP 項グラフの選択を行う．
(5)遺伝操作(複製，逆位，交叉，突然変異)により，次世代の集団を生成する．
(6)終了条件が満たされているときは終了する．そうでなければ(3)へ戻る．

遺伝操作の逆位(TTSP 項グラフの対象とする部分の順序を入れ替える操作)，交叉(2 つの TTSP 項グラフの対象とする部分を入れ替える操作)，突然変異の add-TTSP(TTSP 項グラフにランダムに生成した TTSP 項グラフを加える操作)の適用例を図 3，4，5 に示す．

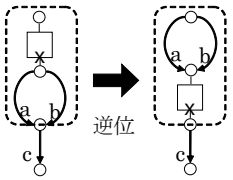


図 3 逆位の適用例

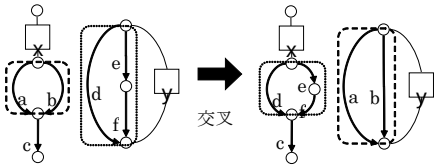


図 4 交叉の適用例

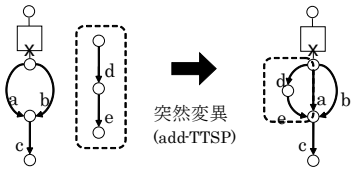


図 5 突然変異(add-TTSP)の適用例

表 1 GP のパラメータ

個体数	50
複製確率	0.05
逆位確率	0.05
交叉確率	0.45
突然変異確率	0.45
最大世代数	200

４．研究成果

本研究の提案手法を実装し，評価実験を行った．実験に用いた GP のパラメータを表 1 に示す．ある TTSP 項グラフにマッチするもの 100 個を正事例，マッチしないもの 100 個を負事例とする．図 7 に正事例と負事例の TTSP グラフの例を示す．正事例と負事例にはそれぞれ 5%のノイズを入れた．

提案手法の 10 試行の平均値を表 2 と図 6 に示す．表 2 は実行時間と最終世代の適合度の最も高い個体(最良個体)のサイズと辺の数および割合を示す．ここで個体のサイズとは，TTSP 項グラフの辺(変数を除く)の総数+変数の総数である．初期個体のサイズをランダムな大きさで生成するようにした．図 6 は各世代の最良個体の適合度の平均値の推移を示す．

図 6 より，提案手法は適合度の高い，特徴的な TTSP 項グラフを獲得することができたといえる．表 2 より，獲得した TTSP 項グラフは具体的あると考えられる．図 8 に獲得した特徴的 TTSP 項グラフの例を示す．図 8 では XXXXXX は変数を表す．

表 2 実行時間と最終世代の最良個体のサイズ

実行時間(s)	612
個体のサイズ	8.6
個体の辺の数	4.1
個体の辺の割合(%)	47

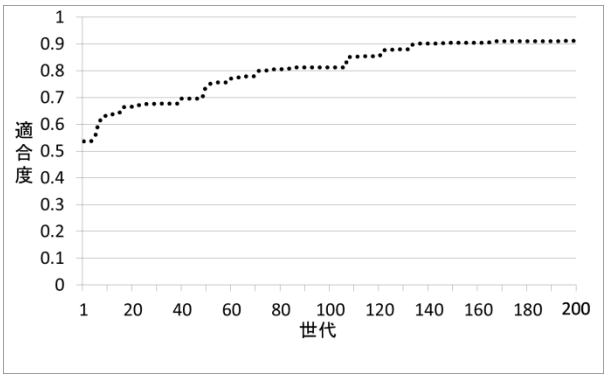


図 6 各世代の最良個体の適合度の推移

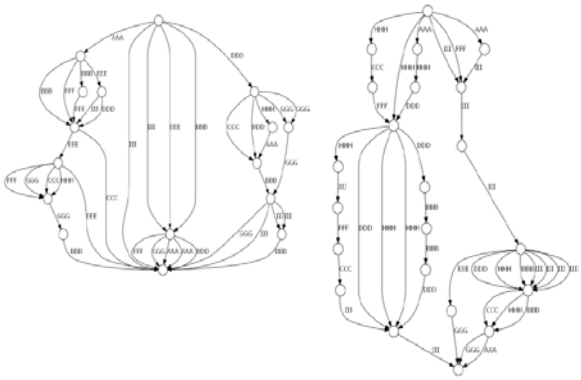


図 7 TTSP グラフの正事例(左)と負事例(右)

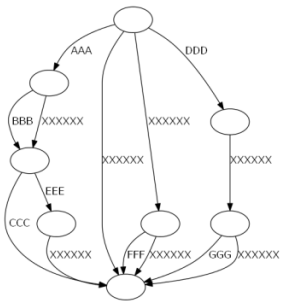


図 8 獲得した特徴的 TTSP 項グラフ

正事例と負事例から特徴的な TTSP 項グラ

フを獲得する機械学習手法以外にも，次の研究成果を得た．木構造データの正事例と負事例から特徴的な木構造パターンを獲得する進化的手法を発展させた．木の編集距離の計算，木構造データに対するカーネル法，グラフパターンの正データからの学習などで成果を得た．

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４件）

① S.Nakai, T.Miyahara, T.Kuboyama, Acquisition of Characteristic Tree Patterns with VLDC's by Genetic Programming, Proc. IADIS Internatinal Conference Information Systems 2013, pp. 332-336.（査読有）

② T.Miyahara, T.Kuboyama, Acquisition of Glycan Motifs using Genetic Programming and Various Fitness Functions, Proc. SCIS-ISIS 2012, pp. 1684-1689.（査読有）
DOI:10.1109/SCIS-ISIS.2012.6505277

③ S.Nagai, T.Miyahara, Y.Suzuki, T.Uchida, Acquisition of Characteristic TTSP Graph Patterns by Genetic Programming, Proc. 2012 IIAI International Conference on Advanced Applied Informatics, pp. 340-344.（査読有）
DOI:10.1109/IIAI-AAI.2012.73

④ Y.Otsuka, T.Miyahara, T.Kuboyama, Learning of Multiple Tree Structured Patterns using Clustering and Evolution, Proc. IADIS Internatinal Conference Information Systems 2011, pp. 227-231.（査読有）

〔学会発表〕（計　３件）

① 中居翔平,宮原哲浩，編集距離と遺伝的プログラミングを利用した特徴的木パターンの獲得，2012 IEEE SMC Hiroshima Chapter 若手研究会，2012年07月04日，広島市立大学

② 永井聡，宮原哲浩，鈴木祐介,内田智之，遺伝的プログラミングによる特徴的なTTSPグラフパターンの獲得，火の国情報シンポジウム2012，2012年3月15日，九州工業大学情報工学部

③ 大塚喜明,宮原哲浩，クラスタリングと遺伝的プログラミングによる複合的木構造パターンの獲得，2011 IEEE SMC Hiroshima Chapter 若手研究会，2011年7月9日,広島市立大学

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

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

(3)連携研究者
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