様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月３日現在

機関番号：１３９０３
研究種目：基盤研究（Ｃ）
研究期間：2008～2010
課題番号：20500132
研究課題名（和文）　ボトムアップ方式による多関係データマイニング手法の確立

研究課題名（英文）　Formulation of a Multi-relational Data-mining Method by Bottom-up Approach

研究代表者
犬塚　信博（INUZUKA NOBUHIRO）
名古屋工業大学・大学院工学研究科・教授
研究者番号：10221780

研究成果の概要（和文）：
本課題の研究目標は多関係データマイニングについて、（１）ボトムアップ手法によるパターンの網羅性と、（２）そのとき効率面でのトレードオフを追求し、（３）化学・創薬の領域や言語・コーパスの領域等の応用的価値を示すことである。さらに（４）理論的観点からグラフマイニング等の他手法との原理的差異を明確化する。

このため本研究で提案する基本パターンとこれに対する演算によって効率よくパターンを網羅するアルゴリズムを設計し、システムを構築する、また適切な応用課題によってシステムの効果を研究した。

研究成果の概要（英文）：
This research project focused on multi-relational data-mining concerning (1) completeness of enumeration of patterns using the bottom-up approach, (2) trade-off between the completeness and efficiency of the procedure and (3) applicability in various areas, such as identification of chemical properties, drug design and analysis of language corpus. Furthermore, we also clarified differences of the approach from other methods including graph mining approach from theoretical and practical aspects.

We designed an efficient algorithm and its variations based on basic properties that we proposed and combination operations for the properties. The algorithm was implemented as a mining system and evaluated in various types of applications

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,000,000	300,000	1,300,000
2009 年度	1,100,000	330,000	1,330,000
2010 年度	1,300,000	390,000	1,390,000
年度			
年度			
総　計	3,400,000	1,020,000	4,020,000

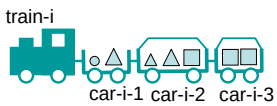
研究分野：知能情報学
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング，論理プログラミング，帰納学習

１．研究開始当初の背景

データベースの１行に１対象がその属性値によって格納されたデータベースで、よく現れる属性値の組合せを枚挙する手法は、頻出項目集合マイニングとして知られ、アソシエーションルールマイニングなどに利用され、多くのマイニング領域で使われる基本手法である。

これを利用するには、データベース全体を１行１エントリーの１関係表になるよう前処理する必要がある。そのため本質的に複数の関係にまたがって定義されるデータでは情報を落とすしかない。項目間のつながりに意味がある場合には有意味なパターンを獲得できない。

たとえば不特定多数個の荷物をもつ貨車を、不特定多数両連ねた列車のデータを関係表にもつとき、このデータに頻出する貨車や荷物のパターンを見つけたい場合（図１参照）を考えると、こうした問題が起こる。有機化合物の構造データも化合物を構成する原子が連結し、その化合物は原子番号やイオン化傾向など各種特徴をもつなど、これと本質的に同じ構造を持ち、毒性など注目する特徴を持つ化合物のクラスからそこに頻出する特徴パターンを取り出すことは直接的応用である。１つの列車や化合物



目標関係表

列車
…
train-i
…

列車	貨車
…	
train-i	car-i-1
train-i	car-i-2
train-i	car-i-3
…	

貨車	長さ	覆い
…		
car-i-1	short	open
car-i-2	long	covered
car-i-3	short	covered
…		

貨車	荷物
…	
car-i-2	load-i-2-1
car-i-2	load-i-2-2
car-i-2	load-i-2-3
…	

荷物	形	大きさ
…		
load-i-2-1	triangle	small
load-i-2-2	triangle	large
load-i-2-3	square	large
…		

図１．複数の関係表が必要な例

に不特定多数の特徴を持つため、１つの表では表現できず、多関係データマイニングという研究領域で議論される。

本研究はこうした極めて一般的であるか困難なデータ領域、すなわち構造的パターンマイニングにおいて利用可能なアルゴリズムを与えるためのものである。

構造的なパターンマイニングではグラフマイニングと帰納論理プログラミング（ＩＬＰ）が有力である。本課題の主軸は後者である。これは計算論理学の理論的資産と Prolog 等の実行系が活用できるため有利である。ＩＬＰは S.Muggleton らが 1990 年に開始した会議で自動プログラミングと機械学習の研究として本格化し、90 年代末からマイニングへ方向を広げた。

提案者はこの領域で実証的機械学習の方法として手法開発を進め、自動プログラミング、並列化による高速化、ファイジー論理プログラミングの拡張、確率的手法との融合、ロボット制御への応用、コーパス解析など、ＩＬＰの手法とその活用を多数の領域で提案してきており[Inuzuka 他 1996, Furusawa 他 1997, Nakano 他 2000, 松井他 2003,Itoh 他 2004, Nakano 他 2006, Motoyama 他 2007 等]、この技術をベースとして構造的データマイニングのアルゴリズム開発を行ってきた。

２．研究の目的

本課題は（１）ボトムアップ手法によって網羅的な頻出パターン抽出にどこまで迫ることができるか、（２）そのとき効率面でおきるトレードオフはどこまで引き上げられるか、（３）化学・創薬の領域と言語・コーパスの領域で応用的価値がどれだけ確保できるのかを示し、（４）理論的観点からグラフマイニング等の他手法との原理的差異を明確化することを目標とした。

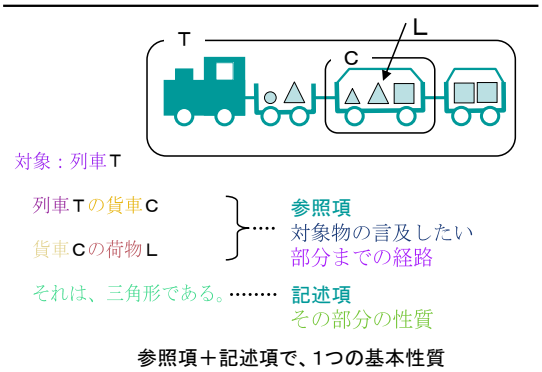


図２．参照項 - 記述項仮説

対象の基本性質を関係述語の論理的性格から規定し、ボトムアップにパターンを構成する手法は他に存在しない。ＩＬＰに基づいたパターン枚挙の方法に限界が指摘されグラフマイニングに研究が集中する中で、本研究は新たな方向を築くことができると考え本課題を設定した。

本アプローチは、マイニング対象の属性に対する基本仮説からスタートしている。対象の基本性質を、その対象からその部分への参照（参照項）と、その部分の性質の記述（記述項）の組からなると仮説する（参照項-記述項仮説：図２）。このアイデアはＩＬＰの従来研究で散見できるが、明確に意識した手法設計に意義がある。

基本性質を定義したことで、多関係ながら、従来の Apriori アルゴリズムを応用でき、見通がよい。ただし、単純な適用では多くのパターンが探索されず、その解決が重要課題であった。

従来のＩＬＰ研究は分類学習などの帰納学習を中心としてきた。論理式はその厳密さに反して、統計的意味に欠け、理論的曖昧さがある。即ち、行っている帰納（一般化）の性格が捕らえにくい。マイニングはパターンの記述を目的とし、一般化は行わない記述学習である。論理式は対象の集合に対応し、問題なのはパターン間の関係や重複の有無などである。先行研究はこれら

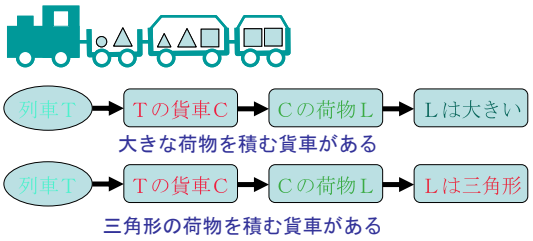


図３．基本パターンの例

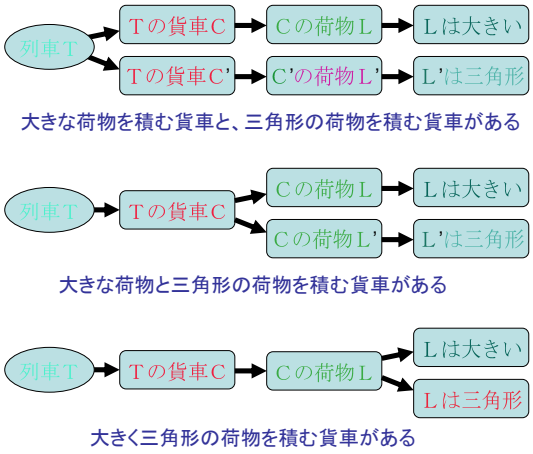


図４．基本パターンの
組合せでできるパターン

の分析に計算論理学が有効であり、アルゴリズム設計に効果的であることを示した。このアプローチは計算論理学の新たな方向を拓く。

３．研究の方法

研究目標に沿って次の項目を進めた。

１．手法の設計とアルゴリズム論的解析

２．手法の実装

３．２つの領域による実験と評価

４．関連分野との理論的考察

先行研究で獲得した関係的属性を基本性質（参照項-記述項仮説）としてボトムアップ的に獲得する基本アイデアを中心に手法を設計し、論理パターンの網羅手法を与えた。

基本性質の具体例を列車の例題を用いて日本語で表現したものを図３に示す。この２つのパターンを基本として組合せることで、多様な性質が得られる。図４は組合せで得られるパターンの例である。関係属性は内部に変数（この場合C,L）を持ち、それらが他の

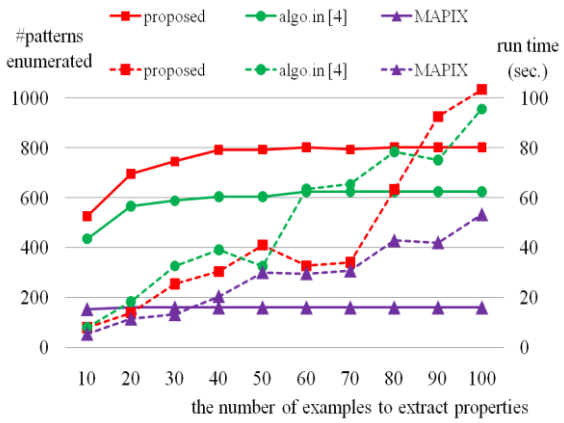


図５．提案法の基本性能の評価

属性の変数と関連する仕方に、多くのパターンを帰着できる。これがアイデアであり、この方法を代数的に記述し、網羅的に探索するアルゴリズムを設計した。

基本性質間の変数によるリンクの管理方法、同値な性質の管理方法（同値な性質も他と組合せたとき、非同値となり捨てられない）を与える。探索空間の制御は、Apriori が用いた束構造の下方閉包性を性質の集合束以外に、包摂関係に関連して多様に利用する手法を設計した。

もう一つのアイデアはボトムアップの徹底であり、基本性質の抽出以外に、変数リンクや基本性質の組合わせもデータへの出現に頼って探索することで効率を得る。

本研究では次の通り研究を進めた。

（１） 上記アイデアを用いたアルゴリズムの設計を行った。

（２） アルゴリズムの正しさを、パターンの性質によって示した。

（３） アルゴリズムを論理型言語によって実装し、人工的データを用いて方法を正しさ、基本的性能を評価した。

（４） 有機化学物質データ、英文構造データを用いて実データによる応用実験を行った。

（５） システムをデータベースと直結するため、ＳＱＬ言語を用いた実装を与え、このシステム上での性能評価実験を行った。

４．研究成果

本研究は我々が従来提案してきたアルゴリズム MAPIX をベースにして、パターンの組みわせを代数的に記述して網羅するアルゴリズムのアイデアを具体化し、アルゴリズムを設計、実装し、人工データを実データによる評価実験を行い、他の研究者による従来法、また、我々の従来法、いずれにも優れた結果を得た。

図５は提案法の基本性能を示す実験の結果である。紫色は我々の従来法であり、基本パターンを用いるが、組合せを用いていない。緑色は我々が本研究を進める前に開発した基本パターンの組合せを用いる手法である。この方法も提案法と目的が同じであるが、代数的組合せを用いていない。赤色が提案法である。□が枚挙されたパターン数、△は計算時間である。提案法は従来法に比べて多くのパターンを枚挙しており、従来法よりも実行時間はかかるが、現実的時間に抑えていることが分かる。このグラフの横軸はパターン枚挙にパターンの種としても用いるサンプルの数であり、これを多数用いると計算時間は要するが、少数のサンプルの時点でパターンの数が飽和しており、用いるサンプル数は少なくてもよいことが示されている。

図６と図７は実データを用いた実験の結果である。図６、図７ともに上図が枚挙されたパターン数、数が実行時間である。横軸は基本パターン生成に用いたサンプル数である。図６は有機化学物質のデータを用いたパターン枚挙の実験である。従来法と比べて提案法は極めて多数のパターンを効果的に得ていることが分かる。このデータは他研究者の従来法ではすべてのパターンを枚挙することに失敗しているもので、本手法の有望さ

を示す。図７は英文の構造データである。データ数が１４００件あり、多数のデータを利

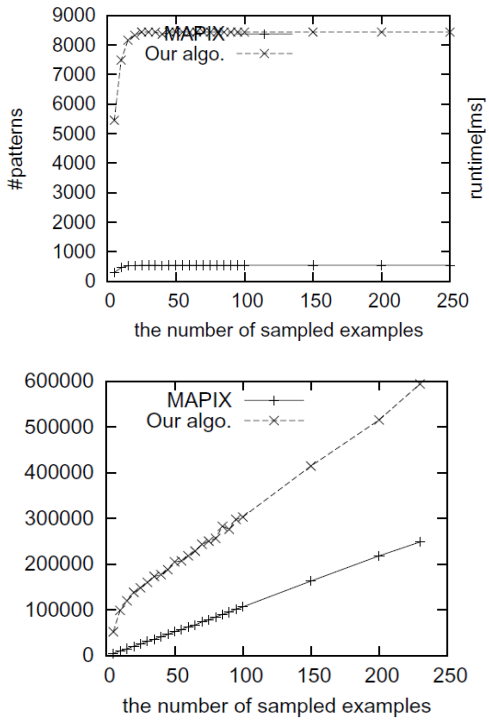


図６．提案法の評価
（有機物質データへの応用結果）

用した場合の実験である。中規模のデータにも対応できることが示された。ただし、このデータにはパターンの構造的複雑さが疎であり、パターンの組合せの効果が少ない。

データベースシステムと組み合わせたシステムの実装においては、大規模なデータへの適用のための実装法として検討した。基本アルゴリズムにおいてはデータベース管理システムとＳＱＬ言語をインタフェースとして結合することに成功し、３倍から１０倍程度の高速の実行が可能であることを示した。しかし、データベースシステムに依存する実装を免れることができておらず、データごとに調整することが必要であり、汎用のシステムとするためには今後の研究が必要である。また、アルゴリズムのすべての機能を実現するには至っていない。

本研究課題を通じて構造的データのパターンマイニングのアプローチとして、ボトムアップ手法が帰納論理プログラミングのアルゴリズムに適しており、効率的なマイニングを実現できることを示した。このことは実データによっても確認された。

本研究は依然としていくつかの重要課題を残している。１つは論理的パターンのマイニングの効果を示しているが、論理式を用いたパターンにおいて、グラフマイニングなど

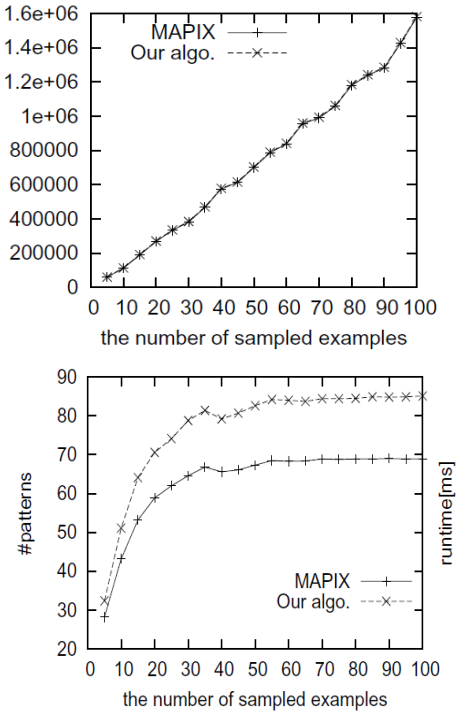


図７．提案法の評価
（英文構造データへの応用結果）

他の手法と比べた場合の問題点が明らかになった。論理式として網羅してもパターとしては不十分な場合がある。この問題は社会ネットワークなどへの応用でも問題となると考えられ、今後の研究が必要である。

またデータベースシステムとの結合は大規模データを用いた実用化のために必要と考えるが、この研究の中で検討した手法には限界があることが分かった。論理手法は必ずしも大規模データを目標としないが、他の実装法も探る必要がある。

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 Nobuhiro Inuzuka, Toshiyuki Makino, “Multi-Relational Pattern Mining System for General Database Systems”, Lecture Notes in Computer Science, Vol. 6278, pp.72-80, 2010, 査読有.

② Nobuhiro Inuzuka, Toshiyuki Makino, “Implementing Multi-relational Mining with Relational Database Systems”, Lecture Notes in Computer Science, Vol.5712, pp.672-680, 2009, 査読有.

③ Nobuhiro Inuzuka, “Relational Pattern Mining Based on Equivalent Classes of Properties”, Lecture Notes in Computer Science, Vol.5012, pp.582-591, 2008, 査読有.

④ Nobuhiro Inuzuka, “Control of Hypothesis Space Using Meta-knowledge in Inductive Learning”, Lecture Notes in Computer Science, Vol. 5078, 911-918, 2008, 査読有.

〔学会発表〕（計６件）

① 中野裕介，犬塚信博，「事例から抽出された特徴に基づく関係型パターンマイニング法と他手法の比較」，第７３回情報処理学会全国大会，2011年3月2-4日，東京工業大学，査読無.

② 中野裕介，犬塚信博，「基本パターンの抽出とその構造的組み合わせに基づくマルチリレーショナル・データマイニング手法」，情報学ワークショップ2010，2010年12月10日，名古屋工業大学，査読無.

③ Yusuke Nakano, Nobuhiro Inuzuka, “Multi-Relational Pattern Mining Based-on Combination of Properties with Preserving Their Structure in Examples”, 20th International Conference on Inductive Logic Programming, 2010年6月27-30日，イタリア フィレンツェ，査読有.

④ Toshiyuki Makino, Nobuhiro Inuzuka, “Implementing Pattern Mining Using Extended Attribute Expression on Relational DB”, Third International Conference on Knowledge Discovery and Data Mining, WKDD 2010, 2010年1月9-10日, タイ プーケット, 査読有.

⑤ Nobuhiro Inuzuka, “Meta-knowledge to explore hypotheses in inductive logic programming”, First Annual Meeting of Asian Association for Algorithms and Computation (AAAC2009), 2009年4月21-22日，中国 香港大学，査読有.

⑥ 牧野敏行，犬塚信博，「関係データベースシステムを結合した論理データマイニングの実装」，情報学ワークショップ2008, 2008年9月25-26日，名古屋大学，査読無．.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等
なし

６．研究組織
(1)研究代表者
犬塚 信博（INUZUKA NOBUHIRO）
名古屋工業大学・大学院工学研究科・教授
研究者番号：10221780

(2)研究分担者
なし

(3)連携研究者
なし