3 グラフベースデータマイニングの基礎と現状

鷲尾　隆（大阪大学産業科学研究所）
washio@ar.sanken.osaka-u.ac.jp

計算機科学や離散構造数学において，グラフ構造は最も研究がなされてきたデータ構造である一方で，構造を持つデータに関するマイニング手法のニーズが増大してきた．このような背景から，近年，グラフベースデータマイニング研究が盛んになってきた．ここではその研究の基礎と現状を紹介する．

■グラフベースデータマイニングの背景

これまでのデータマイニング手法の多くが，表形式やトランザクション形式データを対象としてきた．これに対し，近年，各方面で半構造テキストや順序木，記号系列，グラフ，論理式で表される関係などのデータを対象とする，マイニング手法の研究が盛んになっている．その中で特にグラフは最も一般的な構造の1つである．さまざまな構造データをグラフで表し，そのトポロジーから特徴的関係を発掘するグラフベースデータマイニングの研究が行われている．膨大なグラフから何らかの基準の下で特徴的な部分グラフを発掘するヒューリスティック探索手法として，1990年代半ばにCookとHolderによるSUBDUE[1]と吉田，元田によるGBI[6]が開発された．1998年になってDehaspeとToivonenは，多頻度部分グラフの完全探索を目指すWARMRを発表した[3]．2000年には猪口らがAprioriアルゴリズムとグラフ理論を組み合わせ，非常に高速に完全探索を行うAGMを発表した[4]．

これらの先駆的研究を受け，グラフベースデータマイニングに関する研究論文数は，2002年以降急速に増大しつつある．筆者の把握では，2001年にはSIGMOD，SIGKDD，IJCAI/AAAI，ICML，ECML/PKDD，IEEE ICDMなどの代表的な国際会議で発表されたグラフや木構造データのマイニング手法に関する論文は10件程度であったが，2002年には18件に達した．2003年以降はこれら学会のメイントラック論文に加え，グラフ，木，記号系列構造データマイニングに関するMGTS国際ワークショップに，毎回20件近い投稿があり，この研究分野が拡大期に入ったことをうかがわせる．グラフトポロジーは数学の基本的な研究対象構造であり，また論理言語とも強い関係を持つ．したがって本分野の研究は，データマイニングや機械学習における新しい原理の確立に寄与する可能性が高い．さらに，生物学や化学，材料化学，社会通信ネットワークなどさまざまな実分野において，グラフ構造データは幅広く見られ，実際的応用の高い可能性を有する．本稿では，前半でグラフベースデータマイニングの理論的基礎を概観し，後半で現状の手法の解説を行う．

■グラフベースデータマイニングの基礎

グラフベースデータマイニングは新しい研究分野であるにもかかわらず，多くの原理に依拠している．これは背景のグラフ理論や探索，論理の研究が，すでに豊富な蓄積を持っているためである．この章では，後で述べるグラフベースデータマイニング手法を理解する上で必要となる5つの基礎原理について概説する．スペースの制約上，閉路や並行路を含むラベル付き無向グラフの場合についてのみ記すが，有向グラフやラベルなしグラフについても同様な原理が成り立つ．

部分グラフのクラス

グラフ$G(V, E, f)$は，頂点の集合V，頂点ペアを繋ぐ辺の集合E，辺による頂点の接続関係を表す関数$f : E \rightarrow V \times V$の3項で表される．たとえば，**図-1**(a)に示されるグラフでは，$V=\{v_1, v_2, v_3, v_4, v_5, v_6\}$, $E=\{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9\}$となる．$E$に含まれる各辺$e_h$は，$V$に含まれる$v_i$と$v_j$を$f(e_h)=(v_i, v_j)$によって関連付ける．図の場合には，たとえば$f(e_1)=(v_1, v_2), f(e_2)=(v_1, v_2), f(e_4)=(v_1, v_4), f(e_7)=(v_4, v_4)$となる．

グラフ$G(V, E, f)$の最も一般的な部分構造クラスは，“一般部分グラフ”であり，$V_s \subset V, E_s \subset E$および$e_h \in E_s$であるすべての辺に関して$f(e_h)=(v_i, v_j) \in V_s \times V_s$となる3項関係$G_s(V_s, E_s, f)$で定義される．**図-1**(b)は，元グラフ(a)の頂点v_5および辺e_4, e_6, e_7, e_8, e_9がない一般部分グラフの例である．もう1つの代表的かつ一般的部分構造クラスは“誘導部分グラフ”であり，$V_s \subset V, E_s \subset E$かつ$f(e_h)=(v_i, v_j) \in V_s \times V_s$であるすべての$e_h \in E$について$e_h \in E_s$である3項関係$G_s(V_s, E_s, f)$によって定義される．**図-1**(c)は，元グラフ(a)から頂点v_5を除いた誘導部分グラフの例である．この場合，v_5に直接繋がるe_8とe_9は含まれないが，(b)と異なり元グラフGのv_1, v_3, v_4間に存在するe_4, e_6, e_7は含まれる．3番目の代表的かつ一般的な部分構造クラスは“連結部分グラフ”であり，一般部分グラフでE_s内の辺を辿ってV_s内のすべての頂点間が相互に到達可能となる3項関係$G_s(V_s, E_s, f)$で定義される．**図-1**(d)は，グラフ(c)からさらに孤立頂点v_6を除いた連結部分グラフの例である．

閉路を含むグラフばかりでなく，木やパスも部分グラフのクラスである．閉路を含まないグラフを木という．ここまで頂点や辺のラベルには言及しなかったが，木の辺ラベルを考え，下方(上方)かつ右方(左方)の辺のラベルが常に辞書順により若い値を持つ場合を“順序木”という．また辺に順序がない，ないしはラベル付けされていない木を“非順序木”という．**図-1**(e)は，v_1を根とし辺の番号ラベルで順序付けられた元グラフ(a)の部分順序木である．一方，**図-1**(f)は部分非順序木である．さらに元グラフGの部分構造が分岐を含まない場合を“パス”と呼ぶ．**図-1**(g)は部分パスの例である．

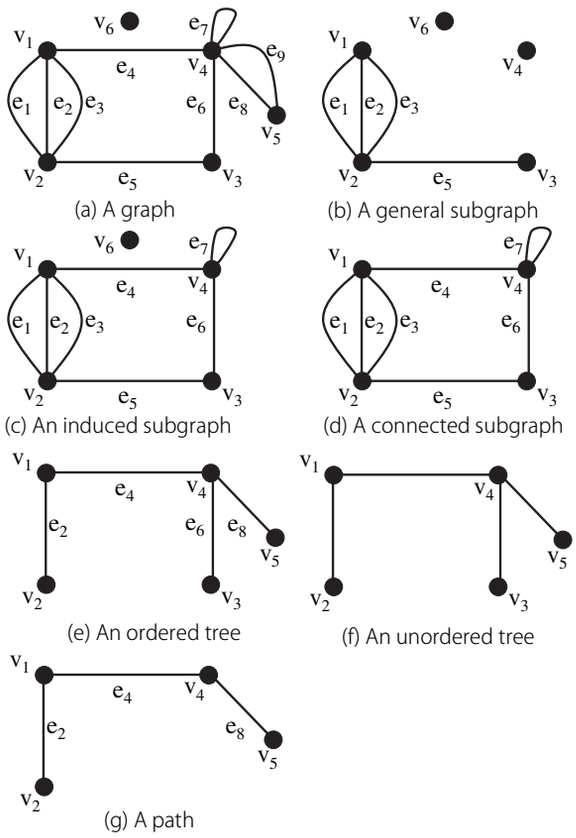


図-1　部分グラフの種類

部分グラフ同型問題

グラフベースデータマイニングにおいては，多数のグラフ間で共通する部分グラフを探索するため，グラフ理論の“部分グラフ同型問題”を拡張した定義を用いる．今，グラフ集合$\mathbf{G}_{\mathrm{all}}=\{G_k, (V_k, E_k, f_k) \mid k=1, ..., n\}$について，以下のようなあるグラフ$G_s(V_s, E_s, f_s)$と$V_s$, V_k間の全単射$g_{sk}(k=1, ..., n)$を見つける問題を，“部分グラフ同型問題”という．

$$e_{kh} \in E_k, f_k(e_{kh}) = (v_{ki}, v_{kj}), \text{ iff } e_{sh} \in E_s, f_s(e_{sh}) = (v_{si}, v_{sj})$$
$$\text{where } v_{ki} = g_{sk}(v_{si}) \text{ and } v_{kj} = g_{sk}(v_{sj}). \quad (1)$$

たとえば，図-1のグラフ(b)と(d)からなる$\mathbf{G}_{\text{all}}=\{(b), (d)\}$について，部分頂点集合$V_s=\{v_{s1}, v_{s2}, v_{s3}\}$と部分辺集合$E_s=\{e_{s1}, e_{s2}, e_{s3}, e_{s5}\}$からなる部分グラフ$G_s(V_s, E_s, f_s)$と全単射$v_{ki}=g_{sk}(v_{si})$ (i=1, 2, 3, $k=b$ or d)は上の条件を満たす．すなわち，(b)と(d)はG_sを共有し，G_sは$\mathbf{G}_{\text{all}}$について部分グラフ同型である．1つの小さなグラフが1つのより大きなグラフの部分かどうかを判定する部分グラフ同型問題は，NP-完全であることが分かっている．上記の複数グラフ間の部分同型問題の複雑性はNP-完全より低いことはあり得ない．

グラフ不変量

同型なグラフは等しい不変量値を持つが，不変量値が等しくても同型なグラフとは限らない．グラフ不変量の例として，グラフに含まれる頂点数や各頂点に接続する辺数(線度)，閉路の数などがある．最も直接的にグラフ構造を表す不変量として，以下の“正準ラベル”がある．同型なグラフは等しい正準ラベルを持ち，正準ラベルが等しければ同型なグラフとなる．

グラフのi-番目の頂点v_iをi-番目の行と列に対応させ，要素によって頂点間の辺の接続関係を表す行列を“隣接行列”という[4]．i, j-要素は，辺の集合$\{e_h\}$で表される．これは厳密には要素が数ではないので行列ではないが，都合上行列と呼ぶ．頂点v_iとv_j間に辺が存在しない場合，要素は空ないし0とする．以下は図-1(a)の隣接行列である．

	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	$\{e_2, e_3, e_4\}$	0	$\{e_1\}$	0	0
v_2	$\{e_2, e_3, e_4\}$	0	$\{e_5\}$	0	0	0
v_3	0	$\{e_5\}$	0	$\{e_6\}$	0	0
v_4	$\{e_1\}$	0	$\{e_6\}$	$\{e_7\}$	$\{e_8, e_9\}$	0
v_5	0	0	0	$\{e_8, e_9\}$	0	0
v_6	0	0	0	0	0	0

無向グラフに関して，その$n \times n$隣接行列から以下のようなコード表現を定義する．

$$x_1, {}_1x_1, {}_2x_2, {}_2x_1, {}_3x_2, {}_3x_3, {}_3 \ldots x_{n-2}, {}_n x_{n-1}, {}_n x_n, {}_n,$$

無向グラフの隣接行列の対角対称性より，上三角部分の要素のみで表される．同様なコードを有向グラフにも定義できる．各要素にその内容の辞書順に番号を振ったとき，あるグラフに一意に対応する正準ラベルは，コード上の要素の並びで番号を並べて得られる数字が最小(あるいは最大)のコードとして定義される．そして，そのコードに対応する隣接行列を“正準形”と呼ぶ．正準ラベルと正準形によって，グラフ表現の多様性や同型問題の探索空間は著しく削減される．

マイニング基準

マイニングの目的や手法上の制約によって，さまざまな基準が用いられる．グラフベースデータマイニングにおいて最もよく用いられる基準は，バスケット分析でも用いられる“サポート”である．グラフデータセットをDとしたとき，それに含まれる部分グラフG_sのサポートは，

$$sup(G_s) = \frac{D\text{内で}G_s\text{が含まれるグラフ数}}{D\text{内のグラフ数}}$$

である．この基準は逆単調性を有し，G_{sy}がG_{sx}の部分グラフであるならば，$sup(G_{sx}) \leq sup(G_{sy})$である．最小サポート“$minsup$”という閾値を設定し，$D$内でそれ以上のサポート値を有する多頻度グラフ$G_s$をすべて導くマイニングが行われる．また最小サポート以上かつ最大サポート以下の値を有する部分グラフを求めるなど，複数の単調性，逆単調性を有する基準を組み合わせたマイニングも行われる．さらに情報エントロピーや情報量ゲイン，ジニインデックス，最小記述長(MDL)など，機械学習分野の基準もよく用いられる[6), 1)]．

探索手法

先述の部分グラフ同型問題をいかに効率的に解くかが，グラフベースデータマイニングの鍵となる．これらはヒューリスティック探索法と完全探索法に分けられる．草創期の手法は，ヒューリスティックなグリーディー探索法であった[6), 1)]．これはさらに深さ優先探索(DFS)と幅優先探索(BSF)に分類される．DFSはメモリ消費を節約できるため，特に初期の研究でよく用いられたが，探索を途中で打ち切る場合，たまたま探索した部分グラフしか得られない．一方，計算機のメモリ容量の増大によって，近年はBSFが用いられる．BSFでは一定の深さまでの同型な部分グラフをすべて探索可能である．しかし多くの場合には，なおもメモリ節約のためにビーム探索などの近似が行われる[1), 6)]．

完全探索が可能な手法の1つは，“帰納論理プログラ

ミング(ILP)”や“帰納推論データベース”の枠組みを用いるものである．“帰納推論”は，適当な仮説を生成しそれによって観測事実の正当化を試みる．この枠組みは，部分グラフ同型問題やマイニングの目的に関連した背景知識を探索に導入可能である．また，グラフ構造を有する関係は“命題論理”によって記述可能であるが，帰納推論はさらにグラフの頂点や辺のラベルが変数であるような“一階述語論理”の関係も導くことが可能である．帰納推論データベースでは，実体化された一階述語形式で表されたデータ(外延的定義の集合)から，最も汎化された一階述語形式規則の完全かつ健全な集合(内延的定義の集合)が導かれる．ILPも帰納推論データベースも，発見される知識が一般性を持つのみならず，正例と負例から知識を導出できる．ただし，一般に探索空間が膨大であり，現実的時間内に探索が終了しない問題がある．

完全探索が可能な手法の2つ目は，バスケット分析のAprioriアルゴリズムで知られる“完全レベル幅探索”である．グラフベースデータマイニングでは，対象データは通常のトランザクション(アイテムの集合)ではなく，トポロジーを含む頂点集合と辺集合の組である．したがって，完全レベル幅探索アルゴリズムは，頂点と辺の接続関係を扱えるように拡張されている[4), 2)]．グラフデータの探索は，1つの頂点のみからなる大きさ1の多頻度グラフを探すことから始まる．これは各種頂点の個数を数えるだけなので容易である．次に，2つの多頻度グラフを組み合わせることで，大きさ2の多頻度グラフ候補が生成される．そして，グラフデータからサポート値が最小サポート$minsup$以上を持つもののみが多頻度グラフとして選ばれる．このように小さな多頻度グラフを組み合わせ，より大きな多頻度グラフを求めることを繰り返す．より大きな多頻度グラフが見つからなくなると，探索を終了する．最大サポートなどの単調性を持つ基準についても同様に適用可能である．

■グラフベースデータマイニング手法

この章では，以上を基に代表的手法について説明する．グラフベースデータマイニング手法は，グリーディー探索法，帰納論理プログラミング(ILP)および帰納推論データベース法，数学的グラフ理論法の3つに大別される．

グリーディー探索法

1994年頃にグリーディー探索に基づく2つの先駆的な手法が発表された．1つ目の手法は“SUBDUE”と呼ばれ，最小記述長(MDL)原理に基づいて，データ内の各グラフGを最も効率よく圧縮する部分グラフを探索する．ある部分グラフG_sでデータを圧縮した際の記述長(DL)は，G_sの記述長$I(G_s)$とデータ内のG_sに相当する個所をすべて1つの頂点で置き換え圧縮したデータの記述長$I(G|G_s)$の和となる．はじめに，同原理の下で単一頂点からなる部分グラフをすべて見つけ，許されたビーム幅内でDLが小さいものを優先的に記録する．次にビーム内に記録された各部分グラフに1つの頂点を付加して拡張し，それらについてさらにビーム幅内でDLが小さいものを記録する．そして，ビーム幅内で最もDLが小さい部分グラフでデータを圧縮する．この拡張，DL記録，圧縮を繰り返し，より拡張，圧縮してもDLが小さくならない準最適な部分グラフを出力する．探索のバックトラックは行われずビームの最大幅も制限されているため，完全性は保証されない．

もう1つの先駆的手法は，“*Graph Based Induction* (GBI)”である[6)]．GBIはSUBDUEに似て，1つの大きなグラフ中に頻繁に現れる頂点のペアを探し，かつそれを1つの頂点に置き換えることによってグラフが最小になるものを選ぶ．ペアの片方または両方の頂点がすでに縮約されていることもあり得るので，縮約は階層構造を持つ．縮約されたグラフや縮約に使われた頂点ペアからなる部分グラフの大きさの指標としては，人為的な定義を用いる．また，遺伝的アルゴリズムに類似したグリーディーなビーム幅探索によって，大きさが極小になる部分グラフが見つかった段階で探索が停止する．GBIはパターン選択の基準としてグラフの大きさばかりでなく，情報利得や情報利得比などの指標も用いることができる．また最新のGBIは，異なる縮約履歴を持つ大きさが同じ部分グラフ同士が同型であるかどうかを各々の正準ラベルを計算して調べ，冗長な探索を極力回避する工夫が導入されている．

ILPと帰納推論データベース法

一般的な連結グラフデータを対象にして多頻度グラフを完全探索する試みは，1998年に発表されたWARMRが最初であると思われる．これはILPの枠組みとバスケット分析のAprioriアルゴリズムに類似したレベル幅探索を組み合わせたもので，化合物の発ガン性予測問題に適用された．この枠組みでは，データをProgolと呼ばれる一階述語論理による言語で表す．たとえば$atomel$ (C,

$A1, c)$, $bond$ $(C, A1, A2, BT)$は，化合物C内の原子$A1$が炭素cであり，同じくC内でタイプBTの結合により他の原子$A2$に接続していることを表す．WARMRはこの表現上で，頻繁に現れる部分連結グラフ構造を探索する．ILPの汎化機能により，述語の引数が変数化されたパターンも探索できるため，グラフよりも一般的なクラスのパターンを導出できる．しかしながら，構造間の同型判定をθ-subsumptionによって行うことと，部分グラフ同型問題の計算複雑性により，小さな連結部分グラフを探索するだけでも計算量の壁にぶつかってしまう．この問題を軽減するため，同様なレベル幅探索法をとりながらも構造間の同型性判定基準を緩めて高速化したFARMARが開発された．これはWARMRに比して2桁高速なマイニングを可能にしたが，同型性判定基準が弱いために同じ部分グラフを異なる形式で導出してしまう難点がある．またいずれの手法も探索可能な部分グラフは，連結部分グラフに限られる．

帰納推論データベースの枠組みを用いた実用的な研究としては，バージョン空間のレベル幅探索アルゴリズムを用いたMolFeaが挙げられる[2]．バージョン空間とは，集合要素間の一般性関係を表すラティス内の部分空間である．この手法はグラフに含まれるパスに関して，単調および逆単調なマイニング基準によって境界が与えられるバージョン空間内を完全探索することが可能である．境界はグラフデータに含まれる全パスの集合Fについて，以下のように与えられる．

$$min(F) = \{f \in F \mid \neg \exists q \in F : q \preceq f\},$$
$$max(F) = \{f \in F \mid \neg \exists q \in F : f \preceq q\}.$$

ここで$f \preceq q$はfがq以上の一般性を持つことを表す．データ集合Dについて1つの基準cを満たすすべてのパスの集合を$sol(c)$としたとき，極小および極大な境界$S(c)$と$G(c)$は以下のように定義される．

$$S(c) = min(sol(c)) \text{ and } G(c) = max(sol(c)).$$

これら2つの境界は$sol(c)$と以下の関係を持つ．

$$sol(c) = \{f \in F \mid \exists s \in S(c), \exists q \in G(c) : g \preceq f \preceq s\}.$$

すなわち，1つの基準c_iを満たす要素の集合$sol(c_i)$は，2つの境界$S(c_i)$と$G(c_i)$に挟まれたバージョン空間である．これらの境界は，基準c_iの下で先に述べたレベル幅探索を行うことによって計算できる．さらに基準の連言からなる条件$c_1 \wedge \ldots \wedge c_n$に対する要素の集合$sol(c_1 \wedge \ldots \wedge c_n)$も，$S(c_1 \wedge \ldots \wedge c_n)$と$G(c_1 \wedge \ldots \wedge c_n)$に挟まれたバージョン空間であり，Hirshのバージョン空間マージアルゴリズムによって，個別の基準のバージョン空間から容易に計算できる．この枠組みを用いて，MolFeaは単調性および逆単調性を有する個々の基準から構成されるさまざまな連言条件を満たす，バージョン空間を完全探索する．たとえば化合物の分子構造データについて，$(c\text{-}o \preceq f) \wedge \neg(f \preceq c\text{-}o\text{-}s\text{-}c\text{-}o\text{-}s) \wedge (sup(f)) \geq 100)$という条件を指定することで，部分パス$c-o$以上に特殊で部分パス$c\text{-}o\text{-}s\text{-}c\text{-}o\text{-}s$以上に一般的ではなく，かつ出現頻度が100回以上の部分パスfをすべて探索できる．MolFeaは発ガン性に関する化合物データから，発ガン性を有する化合物に一定以上多く含まれかつ発ガン性を有しない化合物に一定以下しか含まれない，重要なパスを見つけることに成功した．

数学的グラフ理論法

数学的グラフ理論に基づく手法は，データから部分グラフを完全探索するものがほとんどであり，その際の基準はサポートが主である．この分野の先駆的研究はAGM (Apriori-based Graph Mining) である[4]．探索の基本原理は，バスケット分析のAprioriアルゴリズムに似ており，1つのグラフを1つのトランザクションと見なす．はじめに1つの頂点のみからなる多頻度部分グラフを探索し，それらにさらに頂点を付加して大きさが1つ大きな多頻度部分グラフ候補を生成し，データを検索して実際に多頻度であるもののみを残す．これを繰り返して逐次的により大きな多頻度部分グラフを完全探索する．この際，アイテム集合を扱うバスケット分析と異なり，連結部分グラフに絞って探索する場合には，追加した頂点と元の多頻度部分グラフの間に必ず辺も付加するが，一般部分グラフを探索する場合には辺を付加しない場合も候補とする．これらの処理を効率的に計算機に実装するために，AGMはグラフの表現として隣接行列を用いている．グラフが含む頂点の個数kをグラフの大きさとし，グラフの隣接行列表現をX_k，そのij-要素をx_{ij}, X_kが表すグラフを$G(X_k)$とする．AGMは頂点ラベルや辺ラベルを持つグラフも扱うことができる．頂点ラベルを$N_p = (p=1,\ldots, \alpha)$，辺ラベルを$L_p = (q=1,\ldots, \beta)$とする．効率的実装のために，これらのラベルには自然数が振られている．AGMは有向および無向の一般部分グラフ，誘導部分グラフ，連結部分グラフ，部分順序木，部分非順序木，部分パスなど，グラフデータからさまざまなクラスの多

頻度な部分構造を探索できる．以下に無向誘導部分グラフの探索法について述べる．実際の実装では，より効率的な処理のために隣接行列をグラフ不変量の節に示したコードで表して処理を行う．またコードで定義されるグラフの正準ラベルや正準形を用いる．

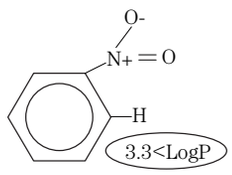
多頻度誘導部分グラフの候補生成は，大きさが1つ小さい2個の多頻度誘導部分グラフを結合することによって行う．大きさがkの2つの多頻度誘導部分グラフを表す隣接行列を，それぞれX_k, Y_kとする．両者がそれぞれk番目の行と列を除いて同一である時のみ，以下のように結合されて大きさ$k+1$の多頻度誘導部分グラフ候補を表す隣接行列Z_{k+1}が生成される．

$$X_k = \begin{pmatrix} X_{k-1} & x_1 \\ x_2^T & 0 \end{pmatrix},\ Y_k = \begin{pmatrix} X_{k-1} & y_1 \\ y_2^T & 0 \end{pmatrix},\ Z_{k+1} = \begin{pmatrix} X_{k-1} & x_1 & y_1 \\ x_2^T & 0 & z_{k,k+1} \\ y_2^T & z_{k+1,k} & 0 \end{pmatrix}$$

ここでX_{k-1}は$G(X_k)$と$G(Y_k)$に共通する大きさ$k-1$のグラフを表す隣接行列であり，$\boldsymbol{x}_i$と$\boldsymbol{y}_i (i=1, 2)$は$(k-1)\times 1$の列ベクトルである．$z_{k,k+1}$と$z_{k+1,k}$の要素は$X_k$と$Y_k$の$k$番目の頂点間の辺ラベルを表す．2つの値は無向グラフの場合には対称性より同一であるが，元のX_k，Y_kからは決まらず2つの場合が考えられる．1つは結合して得られるグラフ$G(Z_{k+1})$のk番目と$k+1$番目の頂点の間にラベルL_qを持つ辺を付加する場合，もう1つはそれらの頂点間に辺を付加しない場合である．これによって$z_{k,k+1}$と$z_{k+1,k}$が“0”か“L_q”である$\beta+1$通りの隣接行列が生成される．X_kとY_kはそれぞれ第1生成行列，第2生成行列と呼ばれる．順序を逆にしてY_kを第1生成行列，X_kを第2生成行列にして結合しても同じ多頻度部分グラフ候補が生成されてしまうため，以下のようにコードがより小さい方を第1生成行列として，冗長な結合を避ける．

$$code(\text{the first matrix}) \leq code(\text{the second matrix})$$

この制約に従って生成された隣接行列を“正規形”という．サポート値の逆単調性から，$G(Z_{k+1})$が多頻度であるためにはその任意の誘導部分グラフも多頻度でなければならない．そこでZ_{k+1}について，大きさが$k \times k$のすべての部分行列が多頻度部分グラフを表すもののみを候補として残す．この確認は前のレベルまでに得られた結果のみを参照すれば行え，元のデータにアクセスする必要はない．こうして得られた候補についてのみ，データにアクセスしてサポート値を計算する．あらかじめ前処理でデータ中のグラフをすべて正規形の行列で表しておくことで，データに対する各候補のマッチングの計算コストを大幅に削減している．以上のすべての処理をレベル幅探索で行い，新たな誘導部分グラフが見つからなくなるまで繰り返す．



変異原性	サポート
高い	13.5%
中程度	20.4%
低い	30.4%
なし	0.0%

図-2　高い変異原性を示す部分分子構造

この手法は230種類の化合物の部分分子構造と変異原性という毒性との間の関係解析に適用された．データの15.2%が高い，45.7%が中程度，9.6%が低い変異原性を持つ．また，原子だけでなく化合物のpHや親疎水性を表すLogPやLUMOなどの物性値も，離散化して付随的な頂点として分子構造に付加された．まずAGMで変異原性が高い化合物に頻出する部分分子構造を抽出し，変異原性が高い化合物でそれを含む割合が，元の分布よりもχ^2-検定で有意に高くなるものを選び出す．これにより**図-2**に示すような高い変異原性をもたらす部分分子構造が発見された．

AGMが提唱されて以降，同様な原理に基づくグラフベースデータマイニング手法がいくつか提案されている．FSG (Frequent SubGraph discovery)は，同様に隣接行列表現と正準ラベルを用いるが，正準ラベル導出の効率を上げるために，コードに頂点の線度などのグラフ不変量を用いる．また，探索対象を連結部分グラフに限り，さらにトランザクションID (TID)法を用いて多頻度部分グラフの生成効率を上げている．FSGでは探索が連結部分グラフに限られ，大規模データではTIDリストの記憶に大量のメモリ空間が必要とされる．より最近ではDFSベースの正準ラベルを用いたgSpan (graph-based Substructure pattern mining)が開発されている[5]．この手法は隣接行列ではなく**図-3**に示されるようなDFS木という表現でグラフを表し，それをコード化して正準ラベルを定義する方法をとっている．グラフ(a)について，どの頂点を木の根とするかによってさまざまな木構造表現が可能である．(b), (c), (d)はグラフの閉路をすべてなくすために必要な点線で表された最小限の辺を除いたも

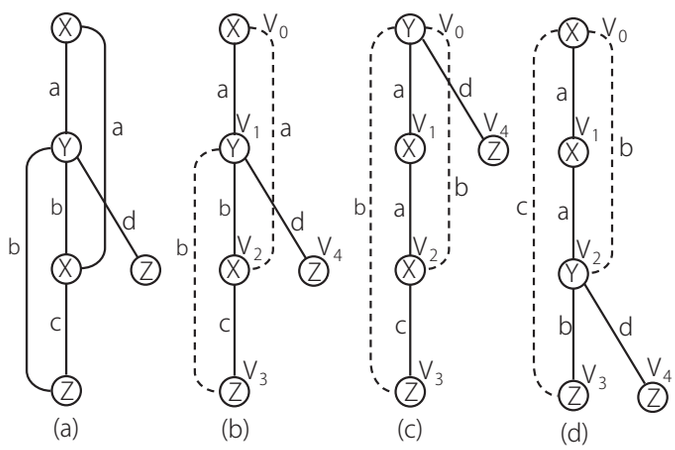


図-3　グラフのDFSコード化

のである．木の根から頂点とその間の辺の組合せに関するラベルをDFSで辞書順に辿り，DFSコードを生成する．(b), (c), (d)はそれぞれ固有のコードを有するが，これらの中で辞書順に見て最小のDFSコードを正準ラベル，およびそれに対応するDFS木を正準形とする．データ内の各グラフはこのようにして複数のコードで表される．そしてデータから得られるすべてのコードを昇順にソートし，コードの第一要素，すなわち最も根に近い要素からソートの順に同じものを1つのノードにマージし，すべてのコードを1つのDFSの順序木にまとめる．このDFS順序木では，左側の枝が右側の枝よりも常にコードが小さい値を持つ．したがってDFS順序木の探索において，あるノードが表す部分グラフがそれ以前に訪問したノードが表す部分グラフと同一である場合には，それ以下の探索はすでに終了しているため枝刈りができる．これによりgSpanは，最小サポート以上の多頻度連結部分グラフの完全探索を，計算量，メモリ消費の両面で効率よく行えるが，非連結を含む一般の多頻度部分グラフを直接探索することはできない．

■今後の展開に向けて

本稿では，部分グラフの種類をはじめとして複数の視点からグラフベースデータマイニングの理論的基礎を説明した．そして，後半で代表的な手法を示した．この分野でさらに研究すべきことは多く，データマイニング手法だけでなく，グラフの特性や同型判定の計算複雑性など理論的にも未解決の問題が多く残されている．この分野は理論，応用の両面で魅力的な研究テーマを多く抱え，データマイニングの重要な研究分野の1つである．

参考文献

1）Cook, J. and Holder, L.: Substructure Discovery Using Minimum Description Length and Background Knowledge, J. Artificial Inte. Research, Vol.1, pp.231-255 (1994).
2）de Raedt, L. and Kramer, S.: The Levelwise version Space Algorithm and Its Application to Molecular Fragment Finding, in IJCAI'01: Seventeenth International Joint Conference on Artificial Intelligence, Vol.2, pp.853-859 (2001).
3）Dehaspe, L. and Toivonen, H.: Discovery of Frequent Datalog Patterns, Data Mining and Knowledge Discovery, Vol.3, No.1, pp.7-36 (1999).
4）Inokuchi, A., Washio, T. and Motoda, H.: Complete Mining of Frequent Patterns from Graphs: Mining Graph Data, Machine Learning, Vol.50, pp.321-354 (2003).
5）Yan, X. and Han, J.: gspan: Graph-based Substructure Pattern Mining, in ICDM'02: 2nd IEEE Conf. Data Mining, pp.721-724 (2002).
6）Yoshida, H., Motoda, K. and Indurkhya, N.: Graph-based Induction as a Unified Learning Framework, J. of Applied Intel., Vol.4, p.297-328 (1994).

（平成16年12月3日受付）