2個組アイテムのデータベースにおける共起成分の含意関係の性質について

Properties of Cofactor Implication for Item Pair Databases

二木 克也* 湊 真一†
Niki Katsuya Minato Shin-ichi

Abstract: ZDD を用いることにより，非明示的に列挙された大規模な組合せ集合データに対して，多様な演算を効率よく実行できると期待されている．本稿では，各レコードが2個組のアイテムで構成されるデータベースにおいて，共起成分の含意関係を抽出した場合の，その結果が持つ意味と，その有用性について考察する．

Keywords: BDD, ZDD, data mining, transaction database, cofactor implication

1 まえがき

近年，大規模記憶装置の発展などによって，大規模なデータベースの中から有用な規則を発見するデータマイニングの研究が盛んになっている．

我々はこれまでに，VLSI CAD の分野で大規模論理関数データの表現法として広く用いられている二分決定グラフ (BDD) [1]，その中でも「ゼロサプレス型 BDD（ZDD: Zero-suppressed BDD） [2]」と呼ばれるデータ構造を用いて，トランザクションデータベースにおける頻出アイテム集合を効率よく生成する手法に関する研究を進めている．ZDD を用いることにより，大規模な組合せ集合データを非明示的に列挙し，頻出アイテム集合の発見から解析に至る多様な演算を効率よく実行することができると期待されている．

一方，ZDD を用いることにより,「共起成分の含意関係」を検出する手法が湊らにより提案されている [3]．共起成分の含意関係が成り立つとは，2つのアイテム a，b に関して，a と共起する成分の集合と，b と共起する成分の集合との間に含意関係が存在することをいう．共起成分の含意関係を抽出し，直接的な含意関係と比較することにより，アイテム同士に直接の含意関係がないにも関わらず，共起成分には含意関係があるようなアイテムの組を検出することができる．これらのアイテムの組には，これまで発見されていなかった，何らかの興味深い関係が存在している可能性がある．

*北海道大学大学院情報科学研究科アルゴリズム研究室, 060-0814 札幌市北区北 14 条西 9 丁目, e-mail niki@mx-alg.ist.hokudai.ac.jp Algorithm Laboratory, Graduate School of Information Science and Technology, Hokkaido University , Sapporo, 060-0814 Japan.
†e-mail minato@ist.hokudai.ac.jp

本稿では，アイテム組合せとしては，最も単純で，基本となる2個組アイテムの集合であるデータベースに注目し，このようなデータベースより共起成分の含意関係を抽出した場合について，関係が抽出される仕組みを議論し，抽出された関係にどのような意味があるのか考察することで，より複雑なデータベースにも通用する一般的な理解を得ることを目的とする．

2 共起成分の含意関係と ZDD

2.1 ZDD

BDD は，図 1 に示すような論理関数のグラフによる表現である．一般に，論理関数のそれぞれの変数について，0，1 の値を代入した結果を，二分岐の枝（0-枝/1-枝）で場合分けし，得られる論理関数の値を，2 値の定数節点（0-終端節点/1-終端節点）で表現すると，図 1 のような二分木状のグラフになる．このとき，場合分けする変数の順序を固定し，冗長な節点の削除と等価な節点の共有という 2 つの縮約規則を可能な限り適用することにより,「規約」な形が得られ，論理関数をコンパクト且つ一意に表せることが知られている．

BDD は元々は論理関数を表現するために考案されたものだが，これを用いて組み合わせ集合データを表現・操作することも出来る．組合せ集合とは,「n 個のアイテムから任意個を選ぶ組み合わせ」を要素とする集合である．これを BDD で表現するとき，類似する組合せが多ければ，部分的に共通する組合せがグラフ上で共有されて，記憶量や計算時間が大幅に削減される場合がある．さらに，組合せ集合に特化した「ゼロサプレス型 BDD

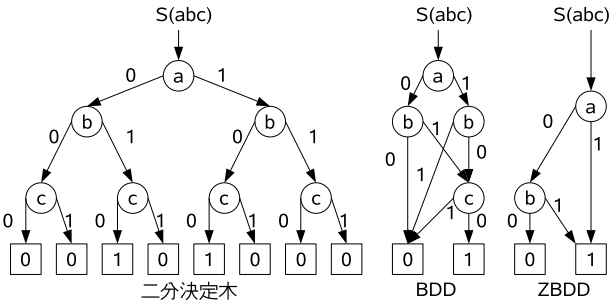

図 1: 二分決定木と BDD，ZDD

（ZDD）」 [2] を用いると，より簡潔な表現が得られ，一層効率よく扱うことが出来る．

ZDD では，冗長な節点を削除する簡約化規則が通常の BDD と異なり，1-枝が 0-終端節点を直接指している節点を取り除く（図 2），という規則になっている．これにより ZDD では図 1 のように，組合せ集合に一度も選ばれることのないアイテムに関する節点が自動的に削除されることになり，BDD よりも効率よく組合せ集合を表現・操作することが出来る．

2.2 共起成分の含意関係

共起成分の含意関係 [3] が成り立つとは，2つのアイテム a, b に関して，a と共起する成分の集合と，b と共起する成分の集合との間に含意関係が存在することをいう．

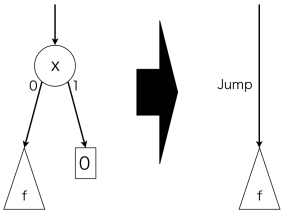

図 2: ZDD の簡約化規則

データベースに含まれる全てのアイテムの組合せの集合をあるアイテム a を含むか含まないか，および，アイテム b を含むか含まないかで $S_{\bar{a}\bar{b}}$, $S_{\bar{a}b}$, $S_{a\bar{b}}$, S_{ab} の 4 通りに分類する（図 3）．このうち，a のみと共起する成分の集合 $(S_{a\bar{b}})$ が，b のみと共起する成分の集合 $(S_{\bar{a}b})$ に含まれるとき，すなわち，$S_{\bar{a}b} \supseteq S_{a\bar{b}}$ が成り立つとき，アイテム a はアイテム b に対して共起成分の含意関係にあると（以下，$a \rightarrow b$ のように記述）いうことができる．また，このとき，$b \rightarrow c$ が成り立つならば，$a \rightarrow c$ も成り立つ．一方，$S_{a\bar{b}}$ が空集合であるとき，a そのものが必ず b と共起するので，すなわち，a は b に対して直接的含意関係にあるということができる．

例えば，以下のような組合せ集合があれば，

$S = \{\{b,c\},\{a,c,d\},\{c,e\},\{a,c\},$
$\{a,b,e\},\{b,d\},\{b,c,d\}\}$

次のような集合へ分けられる．

$S_{\bar{a}\bar{b}} = \{\{c,e\}\}$
$S_{\bar{a}b} = \{\{c\},\{d\},\{c,d\}\}$
$S_{a\bar{b}} = \{\{c,d\},\{c\}\}$
$S_{ab} = \{\{e\}\}$

この例では，$S_{\bar{a}b} \supseteq S_{a\bar{b}}$ が成り立つので，$a \rightarrow b$ が成り立つ，一方，$S_{a\bar{b}} \supseteq S_{\bar{a}b}$ は成り立たないので，$b \rightarrow$a は成り立たない．

共起成分の含意関係を抽出し，直接的な含意関係と比較することにより，アイテム同士に直接の含意関係がないにも関わらず，共起成分には含意関係があるようなアイテムの組を検出することができる．既存の研究の多くは，直接的含意関係に基づいているのに対して，共起成分の含意関係に基づいた研究は少ないので，これらのアイテムの組には，これまで発見されていなかった，何らかの興味深い関係が存在している可能性がある．直接的含意関係とは，図 3 でいうところの $S_{\bar{a}\bar{b}}$，もしくは，S_{ab} のどちらか一方が空集合である状態なので，これを除外し，特に言及のない限り，本稿では，$S_{\bar{a}\bar{b}}$ と S_{ab} がどちらも空集合ではない場合のみを指して，共起成分の含意関係として，取り扱うこととする．

また，抽出された関係について，グラフ化すると，いくつかのクラスタを形成する場合があることがわかっている．このクラスタに関するデータ構造には，何かしらの有意な情報が含まれている可能性が高い．

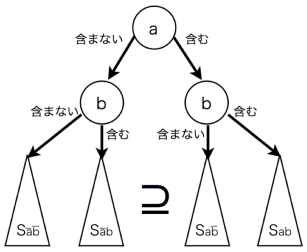

図 3: 共起成分の含意関係

なお，n 個のアイテムが存在するならば，共起成分の含意関係を抽出するためには，$n(n-1)$ 通りのアイテムのペアから制約条件を満たすものを全て抽出する必要があるが，ZDD を用いることにより，データベース中の全てのアイテムのペアに対して，約 n 倍高速なアルゴリズムで検出することが可能である [3]．また，そのようなアルゴリズムを実装したプログラムとして，VSOP [4] が公開されている．

これまでの実験結果としては，現実のインフルエンザのアミノ酸配列のデータベースから，全ての共起成分の含意関係を，現実的な時間で抽出することに成功している [5]．

2.3 2個組アイテムのデータベース

本稿では，2 個のアイテムの組合せが 1 つ 1 つのレコードとなっている

$(\{a,b\},\{a,c\},\{c,d\},\{d,e\},\{b,e\})$

のようなデータベースを対象とする．このようなデータ構造を持つデータベースは日常に多く存在する上に，各アイテムの共起成分がアイテムの組合せの集合ではなく，単独のアイテムの集合であるなど，解析が比較的容易であり，また，アイテムの組合せとして，基本であると思われることから，より複雑なデータベースにおいても通用する，一般論としての共起成分の含意関係が持つ意味を明らかにしていく上で，重要であると考えられる。なお，本稿では，問題を単純化するために，１つのデータベースに同じアイテム組合せが複数回出現するときの頻度や，$\{a,b\}$ と $\{b,a\}$ のようなアイテムの順序は考慮しないものとすることで，データベースを組合せ集合として扱うことにする

3　小規模な例題に関する考察

小規模な例題に対して，VSOP を用いて，共起成分の含意関係を抽出し，Graphviz を用いて，グラフを描画して、考察を加えた．

3.1　核となるアイテムを持つデータベース

実際のデータベースでは，各アイテム間に何かしらの関係が内包されていることが多い．また，データマイニングがその関係を見つけることを目的とする以上，ランダムなデータを想定することは意味がないので，特徴的なデータを与えて試行する．そこで，複数のアイテムの関係の結節点，つまり，複数のレコードに共通して出現するアイテムに着目し，これを核となるアイテムと呼ぶことにする．

3.1.1　核を１つだけ持つデータベース

全ての組み合わせに共通のアイテムが含まれるデータベース

$S = \{\{a,a1\},\{a,a2\},\{a,a3\},\{a,a4\},\{a,a5\},$
$\{a,a6\},\{a,a7\},\{a,a8\}\}$

において，共起成分の含意関係を抽出すると，図 4 のクラスタ構造のように，共通の成分 a を除く全てのアイテム間で双方向の結合が抽出される．これは，a 以外のすべてのアイテム $a1$〜$a8$ が共通の共起成分 a を持っていることから，抽出された関係である．つまり，本稿のように，２個組のアイテムが１件のレコードであるデータベースでは，複数のレコード間で共通のアイテムが含まれているときに，共通ではない方のアイテムの間に関係が抽出される．

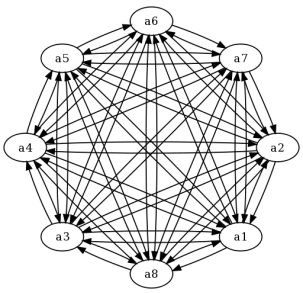

図 4:　１つの核を持つデータベース

3.1.2　核を２つ持つデータベース

前項のデータベースと同様なデータ構造が２つ独立して含まれるデータベース

$S = \{\{a,a1\},\{a,a2\},\{a,a3\},\{a,a4\},\{a,a5\},\{a,a6\},$
$\{a,a7\},\{a,a8\},\{b,b1\},\{b,b2\},\{b,b3\},\{b,b4\},$
$\{b,b5\},\{b,b6\},\{b,b7\},\{b,b8\}\}$

において，共起成分の含意関係を抽出すると，前項で抽出されたものと同じクラスタ構造が２つ抽出された．本稿では，対象とするデータベースを２個組のアイテムペアの集合に限定しているため，共通のアイテムのないアイテム間の関係が抽出されることはない．

次に，アイテム $a1$ と $b1$ を共通の $c1$ に置き換えてみよう．

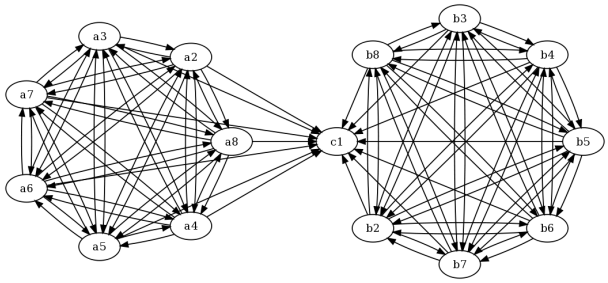

図 5:　２つの核を持ち共通アイテムがあるデータベース

$a1$，$b1$ を指していた矢印が $c1$ でも維持された一方、逆方向の矢印は失われた (図 5)．これは，双方向の矢印が２アイテムの共起成分が完全に一致していることを表していたのに対し，$c1$ の共起成分には，a と b が含まれるため，他のアイテムより，共起成分の数が多くなったためである．

3.2　共起成分の含意関係における推移律

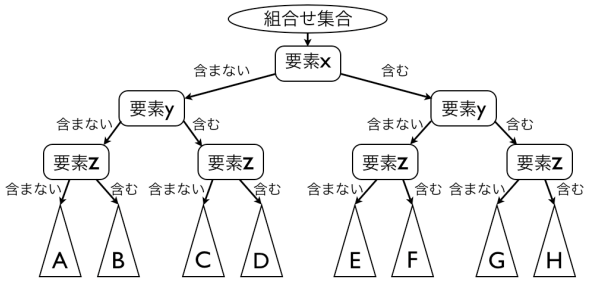

図 6: 共起成分の含意関係と推移律

任意の組合せ集合について，図 6 のように，全ての組合せを，３つの要素を含むか含まないかで分類すると，

８つの集合に分けることができる．なお，３要素を含む場合，これらの要素を組合せより取り除くものとする．また，２個組アイテムの集合であるデータベースでは，集合 I は常に空集合である．

さて，x が y の共起成分の含意関係であるとは，

$(C \cup z \cdot D) \subseteq (E \cup z \cdot F)$

すなわち，

$(C \subseteq E) \wedge (D \subseteq F)$

が成り立っている状態である．

同様に，y が z の共起成分の含意関係であるとは，

$(B \cup x \cdot F) \subseteq (C \cup x \cdot G)$

すなわち，

$(B \subseteq C) \wedge (F \subseteq G)$

が成り立っている状態である．

このことから，

$B \subseteq C \subseteq E \Rightarrow B \subseteq E,\ D \subseteq F \subseteq G \Rightarrow D \subseteq G$

であるので，

$(B \cup y \cdot D) \subseteq (E \cup y \cdot G)$

が成り立つので，共起成分の含意関係において，推移律が成り立つことがわかる．

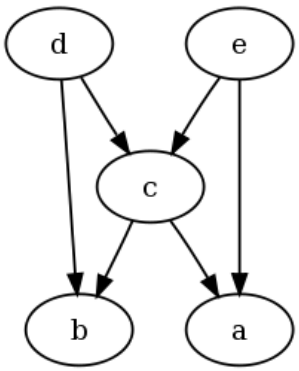

図 7: 共起成分の含意関係と推移律

したがって，共起成分の含意関係は，アイテム同士の順序関係を構成するので，図 8 のように，共起成分の含意関係をグラフ描画する際に，順序関係に沿ってアイテムをレベル付けして配置すると見やすくなる場合がある．なお，ここまでの話において，組合せ集合が 2 個組であるか否かを問題としていないことから，明らかなように，3 個組以上の一般の組合せ集合に対しても推移律は成り立つ．

また，本稿では，$(B \cup y \cdot D)$ が空集合である直接的含意関係を共起成分の含意関係から除外し，議論を展開しているので，$x \to y$,$y \to z$ が共起成分の含意関係として抽出できていても，$x \to z$ が抽出できないことがある．このように，直接的含意関係を除外して考えると，推移律が成立していないように見えるが，このとき，$x \to z$ は直接的含意関係であるので，これを除外しなければ，推移律が成立していることがわかる．

例えば，

$S = \{\{a,b\},\{a,c\},\{a,d\},\{b,c\},\{b,e\}\}$

という組合せ集合において，共起成分の含意関係を抽出すると，図 7 のようなグラフが得られる．a,c,e に関して，a は c の，c は e の共起成分の含意関係にあるが，

$B = \{b\}, C = \{b\}, E = \{b,d\}, G = \{1\}$(他は空集合)

であるから，

$(B \cup c \cdot D) \subseteq (E \cup c \cdot G)$

が成り立ち，B, D も空集合ではないので，a は e の共起成分の含意関係として，関係が抽出されている．

一方，a, c, d に関しては，a は c の，c は d の共起成分の含意関係にあるが，

$A = \{b,d\}, C = \{b\}, E = \{b\}, F = \{1\}, G = \{1\}$(他は空集合)

なので，

$(B \cup c \cdot D) \subseteq (E \cup c \cdot G)$

は成り立つが，B, D は空集合なので，左辺は空集合となり，a は e の直接的含意関係になるので，関係は抽出されない．

また，任意のある関係においてクラスタが分裂するためには，その関係より上位にある全てのアイテムに対して，その関係以下のアイテムが持たない共起成分が存在しなければならない．

$S = \{\{a,b\},\{a,c\},\{a,d\},\{b,c\},\{b,e\},\{d,f\},$
$\{d,g\},\{d,h\},\{f,g\}\}$

という組合せ集合について，抽出された関係を共起成分が多い方が下にくるようにグラフを描くと，図 8 のようになる．ただし，f と g は互いに双方向の関係を持つので，同じ順位を持つことから，１つのノードにまとめている．このとき，例えば，a と h の間において，１つの大きなクラスタを２つのクラスタに分けるためには，h が，a が持っていない，共起成分を持つ必要がある．もし，$\{h,i\}$ という組合せが追加されたならば，図 9 のように分離される．なお，同時に h は，f, g の持っていない共起成分を獲得したことになるので，関係が抽出されなくなっている．

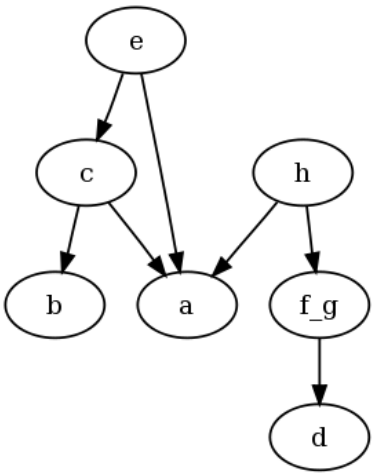

図 8: 共起成分の含意関係をレベル順に表現

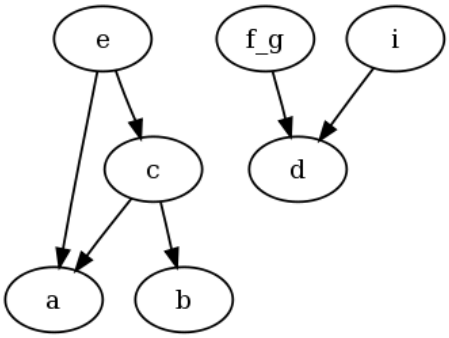

図 9: 図 8 のクラスタを 2 つのクラスタへ分離

3.3 択一型のデータ構造を持つデータベース

アイテムペアの片方が特定のジャンルの情報で，他方が別のジャンルの情報であるというようなレコードのみから構成された，データベースにおいては，ジャンルを跨ぐアイテムがない限り，択一型のデータベースとみなせることから，我々が前回明らかにしたように，ジャンルを跨ぐクラスタは検出されない [6].

3.4 連環状データ構造を持つデータベース

$S = \{\{a_1, a_2\}, \{a_2, a_3\}, \{a_3, a_4\}, \{a_4, a_5\}, ..., \{a_k, a_1\}\}$ のように，連環状のデータ構造が存在するデータベースにおいては，k が 2、ないし、3 の場合では，全てのアイテムの共起成分がそのアイテム自身を除く全アイテムであることから，全てのアイテム間で双方向の関係が抽出される．また，4 の場合では，1 つおきに共起成分が同一のアイテムが存在することから，対角線上のアイテム同士が双方向の関係で抽出される．一方，k が 5 を超えると，関係を抽出することはできなくなってしまう。

4 まとめ

本稿では，2 個組アイテムのデータベースにおいて，実際のデータベースを構成していると思われる要素を想定した例題より，共起成分の含意関係を抽出し，その関係が抽出されたメカニズムについて，考察を行った．2 個組アイテムのデータベースでは，あるアイテムの共起成分とは，そのアイテムが出現するレコードのそのアイテムではない方のアイテムの集合である．このバリエーションが多いということは，重要なアイテムである可能性が高い．また，アイテム間の関係性を基に複数のクラスタ状に抽出することができることから，有用な情報を抽出できる可能性が高い．ところで，実際のデータベースに，この手法を適用した場合，出現頻度が多いアイテムは多くの矢印が入ってくることになり，重要である可能性が高いと考えられる一方で，出現頻度が 1 ないしは少数のアイテムからは，多くの矢印が出ていくことになる．このような関係は，信頼度が低いデータだったり，極めて薄い関係である場合が多いと考えられることから，状況に応じ，無視することも，検討に値する．

今後，今回の考察を基にして、3 個組アイテムのデータベース等，少数のアイテムの組合せからなるデータベースへ考察を進めていく必要がある。

5 謝辞

ご指導をいただいた北海道大学アルゴリズム研究室 Thomas Zeugman 教授に感謝いたします．またご討論いただいた岩崎玄弥さんを始めとする同研究室の方々に感謝いたします.

参考文献

[1] R. E. Bryant : “Graph-based algorithms for Boolean function manipulation,” IEEE Transactions on Competers, Vol. C-35, No. 8, pp. 677–691（1986）

[2] S. Minato: “Zero-Suppressed BDDs for Set Manipulation in Combinatorial Problems,” In Proc. of 30th ACM/IEEE Design Automation Conference (DAC’93), pp. 272–277（Jun. 1993）

[3] 湊: “「共起成分の含意関係」を満たすアイテム集合のデータマイニング,” 人工知能学会　第６５回人工知能基本問題研究会　資料, SIG-FPAI-A603-10, pp. 53–58（Mar. 2005）

[4] 湊: “VSOP：ゼロサプレス型 BDD に基づく「重み付き積和集合」計算プログラム,” 電子情報通信学会コンピュテーション研究会, 信学技報　Vol. 105, No. 72, COMP2005-10, pp. 31–38（May. 2005）

[5] S.Minato and K.Ito: “Symmetric Item Set Mining Method Using Zero-suppressed BDDs and Aplication to Biological Data,” Trans. of the Japanese Society of Artificial Intelligence, Vol. 22, No. 2, pp. 156–164（Feb. 2007）

[6] 二木, 湊: “共起成分の含意関係に基づくデータベースの実験と考察,” 第 12 回情報論的学習理論ワークショップ (IBIS 2009) テクニカルレポート集, pp.99–1004（October. 2009）