２分決定グラフによるデータマイニングシステム

4AH－6

関 弘隆[1]　正代隆義[1]　丸山 修[2]　宮野 悟[2]
[1]九州大学大学院システム情報科学研究科情報理学専攻
[2]東京大学医科学研究所ヒトゲノム解析センター

1　はじめに

大量のデータの中に潜んでいて気がつかない，役にたつ可能性のある知識を発見する技術のことをデータマイニング (*Data Mining*) と総称する．データマイニングは，背景知識を予め持つことなく，与えられた大量のデータのみから知識を得ることを目的とする．Agrawal ら [1] は，データマイニングを，結合ルール (Association Rules)，分類 (Classification)，逐次パターン (Sequential Patterns)，相似列 (Similar Sequences) の４つに分類した．これらのルールは顧客データからの知識発見に対して高い評価を得ており，実用化・商品化が行われている．さらに，Agrawal ら [2] は，与えられたデータから全ての結合ルールを発見するアルゴリズムを提案し，実験的に評価した．しかし，彼らのアルゴリズムは理論的に効率の良いものではない [4]．また，得られた大量の結合ルールから更に知識を獲得する研究もなされている．

一方，ゲノムデータに代表される自然科学のデータにおいても，結合ルールの発見が試みられている [5] が，自然現象には直観的に予測できる規則性が少ないため，結合ルールでは我々の望む知識の発見が難しい．

本論文では，結合ルールを２部決定グラフ (Binary Decision Diagram (BDD)) により一般化した２部決定グラフ結合ルール [4] を発見する問題について考察する．２部決定グラフを用いることにより，論理積，論理和と否定による知識表現が可能になる．２部決定グラフ結合ルールは，２つの BDD(それぞれ，LBD，UBD と呼ぶ) からなる．多くの場合，LBD はユーザが自然な形で与えることが多いため，本質的にこの問題は UBD を発見する問題に帰着する．本論文では，現実的な時間で UBD を見つけるための条件について考察する．さらに，Quinlan による ID3 アルゴリズム [6] を応用した UBD を発見するアルゴリズム [4] を，ゲノムデータに適用した実験結果について述べる．

Mining Binary Decision Diagram
Hirotaka Seki, Takayoshi Shoudai
Department of Informatics, Kyushu University
Kasuga 816, JAPAN
Osamu Maruyama, Satoru Miyano
Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108, JAPAN

2　２分決定グラフ結合ルール

２つの異なる有限集合 O，L と，$O \times L$ から $\{0,1\}$ への関数 f が与えられたとき，３つ組 $D = (O, L, f)$ をデータ (*data*) と呼ぶ．O に属する元のことをオブジェクト (*object*)，L に属する元のことをラベル (*label*)，もしくは属性 (*attribute*) と呼ぶ．

データ $D = (O, L, f)$ 上の２分決定グラフ (Binary Decision Diagram，BDD) とは論理関数を表現する連結な非循環グラフであり，各頂点は次の３つのいずれかである．(1) 根 (*root*): 入次数 (indegree)2 である唯一の頂点．(2) 内部頂点 (*internal node*): 出次数 (outdegree)2 の頂点．(3) 末端頂点 (*terminal node*): 出次数 0 の頂点．末端頂点のうちの１つを特に指定し目的頂点 (*goal node*) と呼ぶ．根及び内部頂点には L の要素がラベル付けしており，さらにその頂点から出る２つの辺には 0 と 1 がそれぞれ対応づけられている．

ある O の元 s が２分決定グラフ B を充足する (satisfy) とは，s が次の手続きで根から目的頂点に到達するときをいう: まず v を根とする．v が末端頂点になるまで次を行う．v のラベル $\ell(v)$ としたとき，$f(s, \ell(v)) = 1$ (resp. 0) ならば 1 (resp. 0) とラベル付けされた辺の先の頂点に移動し，新しくその頂点を v とする．このとき，

$$S(B) = \{s \in O | s \text{は} B \text{を充足する}\}$$

と定義する．

２つの実数 $0 \leq p_u, p_\ell \leq 1$ が与えられたとき，データ $D = (O, L, f)$ 上の２部決定グラフ B_u, B_ℓ が，次の２つの条件を満たすとき，(B_u, B_ℓ) を (p_u, p_ℓ) に関する２分決定グラフ結合ルールという:

$$(1)\ \frac{|S(B_u)|}{|O|} \geq p_u, \quad (2)\ \frac{|S(B_u) \cap S(B_\ell)|}{|S(B_u)|} \geq p_\ell.$$

このとき，B_u を Upper Binary Diagram (UBD)，B_ℓ を Lower Binary Diagram (LBD) と呼ぶ．

3　UBD 問題

２分決定グラフ結合ルール (B_u, B_ℓ) において，B_u, B_ℓ の２分決定グラフのラベルは，それぞれ全く異なるものでなければならない．このことは言いかえると，ある種類のラベルを持つ２分決定グラフ B_ℓ により説明できるデータを，全く異なる種類のラベルを持つ B_u で説明したものが２

分決定グラフ結合ルール (B_u, B_ℓ) であることを意味する．次の問題を考える．

UPPER BINARY DIAGRAM
入力: データ $D = (O, L, f)$, 2分決定グラフ B_ℓ, 2つの実数 $0 \leq p_u, p_\ell \leq 1$.
問題: (B_u, B_ℓ) が (p_u, p_ℓ) に関する2分決定グラフ結合ルールとなるような2分決定グラフ B_u が存在するか．

この問題に関して，次の結果を得た．

Theorem 1 求めたい2分決定グラフを $kDNF$ $(k \geq 1)$ と同値なものに限っても，**UPPER BINARY DIAGRAM** は NP 完全である．

p_u もしくは p_ℓ のいずれか一方が定数であったとしても，NP 完全である．しかし，p_u と p_ℓ が両方とも定数である場合の同問題の計算量は未解決である．同様に，kCNF $(k \geq 1)$ に関しても，同様の結果，すなわち NP 完全であることを証明できる．

4 実験結果とこれからの課題

前章で述べたように，2分決定グラフを求める問題の多くは NP 完全となる．従って，我々は，Quinlan によって提案された ID3 アルゴリズム [6] と Bryant によって提案された2分決定グラフの簡約アルゴリズム [3] を応用した2分決定グラフ結合ルールを求めるアルゴリズムを使用し，実験を行った (図 1,2)．

求めたい2分決定グラフの幅や深さを制限した場合，自明に多項式時間アルゴリズムが得られることがある．しかし、より豊かな知識表現を記述できる2分決定グラフを求めるためには，その形，2分決定グラフ結合ルールの条件について議論する必要がある．

参考文献

[1] R. Agrawal, T. Imielinski, and A. Swami, Mining Association Rules between Sets of Items in Large Databases, *In Proseedings of the 1993 International Conference on Management of Data(SIGMOD 93)*, pp. 207–216, 1993.

[2] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, Technical Report RJ9839, IBM Almaden Research Center, 1994.

[3] R. E. Bryant, Graph-Based Algorithms for Boolean Function Manipulation, *Transactions on Computers*, Vol. 35, No. 8, pp. 677–691, 1986.

[4] 井上和哉，2分ダイアグラム結合ルールのデータマイニング，九州大学大学院システム情報科学研究科修士論文, 1996.

[5] G. Shibayama, K. Satou, and T. Takagi, mining Association Rules from Signals found in Mammalian Promoter Sequences, *Proc. the 6th Genome Informatics Workshop*, pp. 108–109, 1995.

[6] J. R. Quinlan, Induction of Decision Trees, *Machine Learning*, Vol. 1, pp. 81–106, 1986.

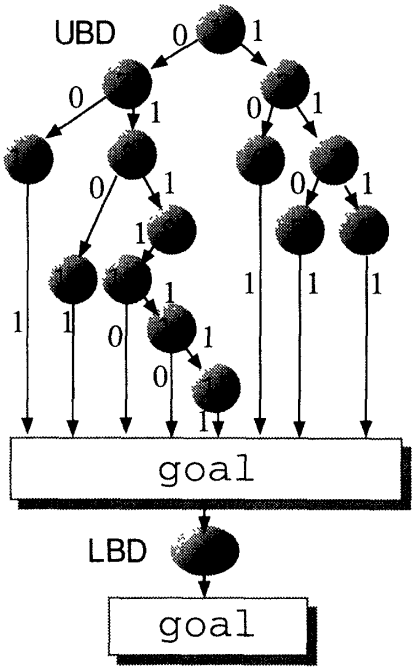

図 1: オブジェクトの集合は，E-coli(大腸菌)DNA のタンパク質コーディング領域 (CDS)(4285 個) である．目的頂点以外の末端頂点は省略した．LBD は，`Physiological Categories` における `Cell processes` の `Cell Division`，さらに `Cell Division` に該当する CDS(29 本) のみを選択する．(`http://DBASE.mbl.edu/ecoli/menu.htf`)

1 H : cggactttag
2 T : cgacaggcac
3 H : gggaagccag
4 (T : ttgccgcttt) ∧ ($\overline{\text{B : ASVY}}$)
5 (B : VVQD) ∧ (T : aggcgaaaga) ∧ ($\overline{\text{B : DEAA}}$) ∧($\overline{\text{B : AADS}}$)
6 (B : RDQE) ∧ ($\overline{\text{B : LSAL}}$) ∧ ($\overline{\text{B : VPPW}}$) ∧(T : tcattggcgt) ∧ ($\overline{\text{B : AAFV}}$)
7 H : aagtactatt
8 $\overline{\text{B : AVDA}}$
9 B : VDEF
10 T : tcaatagaga
11 ($\overline{\text{B : DEFE}}$) ∧ (B : AAAA)
12 (B : MQMK) ∧ (B : ADDI)
13 (B : HEND) ∧ (T : aaggtgccgc)
14 (B : HMME) ∧ (T : ccctaagcac) ∧ ($\overline{\text{B : AADD}}$)

図 2: 図 1 の 2 分決定グラフの頂点のラベル．ある CDS に対して，それより上流 300 文字をヘッド，下流 200 文字をという．H, B, T はそれぞれ上流，CDS，下流を示し，各文字列は，該当する領域にその文字列が存在するか否かを表す．