On the Complexity of Data Mining for Binary Decision Diagram Rules

Osamu Maruyama[†] Takayoshi Shoudai[*] and Satoru Miyano[†]
丸山修 正代隆義 宮野悟

[†]Human Genome Center, Institute of Medical Science, University of Tokyo

[*]Department of Informatics, Kyushu University

{maruyama,miyano}@ims.u-tokyo.ac.jp, shoudai@i.kyushu-u.ac.jp

要旨

We discuss the problem of data mining for binary decision diagram rules (BDDRs). We show that the problem is, in general, NP-complete, and report some results of a preliminary experiment on biological databases with devising a heuristic algorithm of mining BDDRs.

1 Introduction

The term "data mining" has been applied to a broad range of activities that attempt to discover new knowledge from existing data, where usually the original data is a collection of information gathered in a way. The problem of data mining was first formulated by agrawal *et al*, [2, 3]. In the problem, we consider a set X of objects, of which a set of attributes are defined, and are given the values of the attributes for each object in X. The task is to find relationships between various attributes.

The most well-studied problems in data mining is the search for *association rules* on items [2, 3, 4, 14, 15, 13, 11, 22, 23, 16, 24, 7, 8], which are intended to identify rules of the type "A customer purchasing item X also purchases item Y" which is supported by a number of transactions gathered in retail stores. Formally, an association rule is an implication of the form $X \to Y$, where X and Y are sets of some items (i.e., objects).

The interestingness of a rule is a key concept in data mining. The interestingness of an association rule is measured via *support* and *confidence*. The support for a rule is defined as the fraction of transactions that satisfy the union of items in the consequent and antecedent of the rule. The confidence for a rule is defined to be the fraction of transactions satisfying X that also satisfy Y. For the association rule $X \rightarrow Y$ to hold, both of the support and confidence of the rule must exceed given support and confidence thresholds.

There would be numerous applications of data mining which fit into this framework. One of the more challenging applications of data mining is the knowledge discovery in Molecular Biology and Genome Informatics. One of the reason is that, in the areas, there are various huge databases, which are still growing up by gathering outputs produced in biological experiments. Another reason is that even a method of tools working on such a database to help biologist's knowledge discovery is not established yet. In [20, 19], the scheme of data mining for association rules is applied to the problem of finding some rules existing in protein databases which include the sequence features, the structural features and the functional features of proteins. They reported identifying some previously unknown correlations between such features.

The goal in our research on data mining is to produce, by applying a data mining method we develop to biological databases, biological knowledge which helps biologists analyze the world of genome. However, association rules as a model of knowledge representation might be crucial to capture various and complex rules, correlations and structures existing in databases, especially genome databases. We then consider, instead of association rules, *binary decision diagram rules* (BDDRs), which is based on BDDs and a generalization of association rules [1]. A BDD is well-known as a model of Boolean functions and has been studied extensively [12, 6, 21]. A reason of that would be that a BDD is visible representation, that is also true in BDDRs. A BDDR is a kind of BDDs which consists of two BDDs, B_a and B_c, where each of B_a and B_c has exactly one specified node, called the *goal node*, and the goal node of B_a is identified with the root node of B_c. The BDDs B_a and B_c play roles of the antecedent and consequent of the rule, respectively. We denote the BDDR by (B_a, B_c).

We in this paper consider two kinds of the interestingness of a BDDR. One of them is defined as follows; For a set X of objects, $X' \subseteq X$ the set of objects reaching the goal node of B_a, and $P \subseteq X$ the set of objects reaching the goal node of B_c. We call $|X'|/|X|$ and $|X' \cap P|/|X'|$ the *I/O ratios of B_a and B_c*, respectively. They are the ratio of the number of the objects reaching the goal node to the number of the input objects in each BDD. The I/O ratio of B_c exactly corresponds to the confidence of an association rule. The value of the product of the I/O ratios of B_a and B_c corresponds to the support of a association rule. We say that a BDDR $r = (B_a, B_c)$ holds with respect to the I/O ratios $0 \leq \varepsilon_a, \varepsilon_c \leq 1$ if the I/O ratios of B_a and B_c are greater than or equal to ε_a and ε_c, respectively. Another interestingness of a BDDR r is $\frac{|X'-P|+|P-X'|}{|X'|}$, which captures the symmetric difference between X' and P. We call it the *symmetric difference ratio of r*.

The problem we address here, which would be one of the basic problems in this framework of data mining for BDDRs, is to find a BDD B_a such that, for a given BDD B_c, the BDDR (B_a, B_c) holds with respect to given thresholds of the I/O ratios. We denote this problem by **ABDD**, an abbreviation of "antecedent of a BDD rule." The method of solving **ABDD** would be useful when, in mining BDDRs, a user fixes the consequences part of BDDRs, when the consequent parts of BDDRs are enumerated (the efficient enumeration of BDDs is another problem in this framework of data mining for BDDRs), and so on. Another problem in data mining for BDDRs, which is to identify the consequent part of a BDDR, is discussed in [1].

We first show that **ABDD** is NP-complete when the antecedent BDD of a BDDR is restricted to a BDD equivalent to a disjunctive normal form consisting only of positive literals even if the number of literals in each clause fixed to a constant $k \geq 1$. The case where "disjunctive normal form" is replaced with "conjunctive normal form" in the above problem is shown to be also NP-complete, which is equivalent to the problem of finding the antecedent of an association rule. The case for conjunctive normal form is NP-complete even if either thresholds is a fixed constant. Although such BDDs have relatively simple forms, the task to mine them seems to be intractable. For the problem where the interestingness of a BDDR is measured by the symmetric difference ratio, we also have the same result concerning with disjunctive normal form.

Finally, we describe a preliminary experiment on data mining for BDDRs in biological databases with a heuristic algorithm.

2 Preliminaries

For two finite sets X and Y, let $f : X \times Y \to \{0, 1\}$. We call an element of X an object, and an element of Y an *attribute* of X. We define *data* as $D = (X, Y, f)$.

Let $D = (X, Y, f)$ be data. A *binary decision diagram (BDD for short) on* Y is a rooted, directed and acyclic graph, each node of which is either a *terminal node* or a *non-terminal node.* A terminal node is of out-degree zero. A non-terminal node, which has two outgoing edges labeled with 0 and 1 respectively, is labeled with an attribute of Y. The *root node* is the node of in-degree zero. Exactly one terminal node is specified as the *goal node of* G. We denote the label of a node v by $l(v)$. We say that *an object* $x \in X$ *satisfies* G if x reaches the goal node of G in the following way; At the root node v, one first test whether $f(x, l(v))$ is equal to zero or not; if $f(x, l(v)) = 0$ then one traverses the edge labeled with 0, otherwise the edge labeled with 1 is traversed. The procedure is repeated at the node one reach until one reaches a terminal node. We denote $S(G) = \{x \in X \mid x \text{ satisfies } G\}$. Let B_a and B_c be BDDs on Y. A *binary decision diagram rule* (BDDR for short), denoted by

(B_a, B_c), is a BDD consisting of B_a and B_c, where the goal node of B_a is identified with the root node of B_c. For a BDDR $r = (B_a, B_c)$, we call B_a and B_c the *antecedent* and *consequent* of r, respectively. We define the I/O ratios of the antecedent B_a and consequent B_c as $\frac{|S(B_a)|}{|X|}$ and $\frac{|S(B_a) \cap S(B_c)|}{|S(B_a)|}$, respectively. We also define the symmetric difference ratio of r as $\frac{|X'-P|+|P-X'|}{|X'|}$.

Let $r = (B_a, B_c)$ be a BDDR. For $0 \leq \varepsilon_a, \varepsilon_c \leq 1$, we say that a BDDR holds with respect to the I/O ratios, ε_a and ε_c if the I/O ratios of B_a and B_c are greater than or equal to ε_a and ε_c, respectively. For $\varepsilon \geq 0$, a BDDR is said to hold with respect to the symmetric difference ratio ε if the symmetric difference ratio of r is less than or equal to ε.

3 Finding the antecedent of a BDDR

When one mines a kind of if-then rules, e.g., association rules and BDDRs, one might hope to specify the consequent part of the rules. We now consider the problem corresponding to such a situation. The problem **ABDD** (antecedent of a BDD rule) is defined as follows; given a BDD B_c and $0 \leq \varepsilon_a, \varepsilon_c \leq 1$, to find a BDD B_a such that the BDDR (B_a, B_c) holds with respect to the I/O ratios ε_a and ε_c.

We first consider the case where the BDD to be the antecedent of a BDDR is restricted to a disjunctive normal form consisting only of positive literals.

Theorem 1 *The problem* **ABDD** *is NP-complete when the antecedent BDD of a BDDR is restricted to a BDD equivalent to a disjunctive normal form consisting only of positive literals even if the number of literals in each clause fixed to a constant* $k \geq 1$.

It is easy to prove Theorem 1 (We gave a reduction from the minimum set cover problem [10]). The case where "disjunctive normal form" is replaced with "conjunctive normal form" in the above problem is also intractable as follows.

Theorem 2 *The problem* **ABDD** *is NP-complete when the antecedent BDD of a BDDR is restricted to a BDD equivalent to a conjunctive normal form consisting only of positive literals even if the number of literals in each clause fixed to a constant* $k \geq 1$.

The case for conjunctive normal form is NP-complete even if either threshold of the I/O ratio of the antecedent or the consequent is a fixed constant.

Next we consider the symmetric difference ratio of a BDDR $r = (B_a, B_c)$. Recall that the symmetric difference ratio of r is $\frac{|S(B_a)-S(B_c)|+|S(B_c)-S(B_a)|}{S(B_a)}$. The problem **ABDD**-SD is defined as follows; given a BDD B_c and $\varepsilon \geq 0$, to find a BDD B_a such that the BDDR (B_a, B_c) holds with respect to the symmetric difference ratio ε.

Theorem 3 *The problem* **ABDD**-*SD is NP-complete when the antecedent BDD of a BDDR is restricted to a BDD equivalent to a disjunctive normal form consisting only of positive literals even if the number of literals in each clause fixed to a constant* $k \geq 1$.

4 Preliminary experiment

Although **ABDD** is intractable, we have devised an ad hoc scheme for solving the problem and executed a preliminary experiment on genome databases in order to probe the feasibility of usefulness of the method of data mining for BDDRs.

In the scheme, first one makes a decision tree T by the ID3 algorithm [17]. Next T is given to Bryant's algorithm [9], which reforms T into an equivalent BDD. Fig. 1 is an antecedent produced by the scheme. The objects in the data mining problem are the coding regions (CDS) of the complete genome sequence of E. coli [5] (See `http://www.genetics.wisc.edu`). The total number of CDSs is 4,285. For a CDS s, we define the *upstream of* s as the substring starting three hundred bases upstream and ending just before s. The *downstream of* s is defined as the substring next to s two hundred bases downstream. The attributes for the antecedent BDD are as follows; all of the substrings with length ten of the upstream and downstream of s, and all of the substrings with length four of the translation (that is the primary sequence of the protein) of s. The label of each node is listed in Fig. 2. For a CDS x, the value of an attribute y is defined as follows; if y is originally a substring of the upstream (downstream) of a CDS and approximately matches a substring of the upstream (downstream, respectively) of x then returns one, zero otherwise; if y is originally a substring of the translation of a CDS and approximately matches a substring of the translation of x then returns one, zero otherwise [25].

The consequent BDD in this execution is satisfied with the CDS whose gene name is included in some class classified in GenProtEc -E. coli Gene and Gene Product Database [18]. That class is Cell Division of Cell Division of Cell processes in Physiological Categories (See `http://www.mbl.edu/html/ecoli.html`).

The meaning of the BDDRs produced in this time is now examined.

参考文献

[1] 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. Master's thesis, 九州大学大学院システム情報科学研究科, 1996.

[2] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, December 1993.

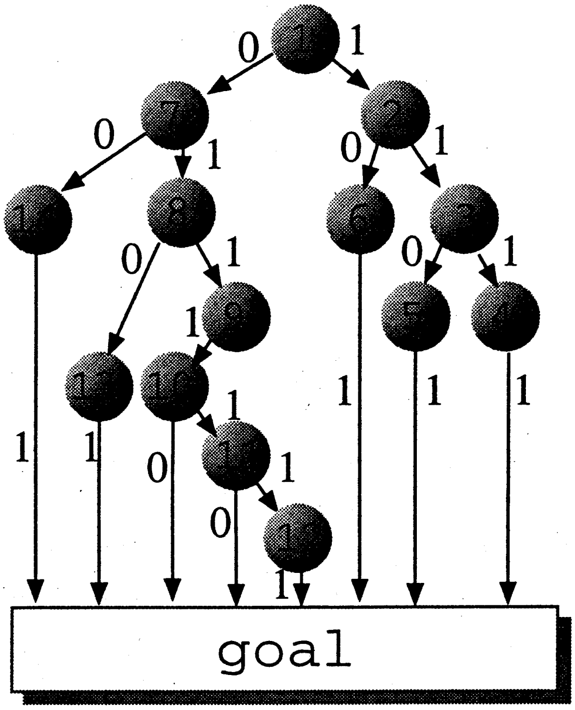

図 1: The BDD is an antecedent produced in the scheme. The terminal nodes except the goal node is omitted. The label of each node is listed in Fig. 2.

[3] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[4] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.

[5] F.R. Blattner, G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, and Y. Shao. The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, 1997.

[6] B. Bollig and I. Wegener. Improving the variable ordering of OBBDs is np-complete. *IEEE transactions on computers*, 45:993–1002, 1996.

[7] S. Brin, R. Motowani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 265–276, 1997.

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

図 2: The list of labels of nodes in Fig. 1. The signs, H, T and B, means upstream, downstream and translation. Some nodes are put together by the operation $\wedge$.

[8] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 255–264, 1997.

[9] R.E. Bryant. Graph-based algorithms for Boolean function manipulation. *IEEE Transactions on Computers*, 35:677–691, 1986.

[10] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman and Company, New York, 1979.

[11] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. of the 21st International Conference on Very Large Data Bases*, pages 420–431, 1995.

[12] K. Hirata, S. Shimozono, and A. Shinohara. On the hardness of approximating the minimum consistent OBDD problem. In *Proc. 5th Scandinavian Workshop on Algorithm Theory*, pages 112–123, 1996.

[13] M. Houtsma and A. Swami. Set-oriented mining for association rules in relational databases. In *Proc. of the International conference on data Engineering*, pages 25–33, 1995.

[14] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. of the 3rd international conference on information and knowledge management*, pages 401–407, 1994.

[15] H. Mannila, H. Toivonen, and A.I. Verkamo. Efficient algorithms for discovering association rules. In *Proc. of the AAAI workshop on knowledge discovery in databases*, pages 144–155, 1994.

[16] J.S. Park, M.S. Chen, and P.S. Yu. An effective hash based algorithm for mining association rules. In *Proc. of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 175–186, 1995.

[17] J. R. Quinlan. Induction of decision trees. *Machine*, 1:81–106, 1986.

[18] M. Riley and B. Labedan. E. coli gene products: Physiological functions and common ancestries. In F. Neidhardt, R. Curtiss, III, E.C.C. Lin, J. Ingraham, K. B. Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger, editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd*, pages 2118–2202, Washington, D.C., 1996. ASM Press.

[19] K. Satou, T. Ono, Y. Yamamura, E. Furuichi, S. Kuhara, and T. Takagi. Extraction of substructures of proteins essential to their biological functions by a data mining technique. In *Proc. of Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 254–257, 1997.

[20] K. Satou, G. Shibayama, T. Ono, Y. Yamamura, E. Furuichi, S. Kuhara, and T. Takagi. Finding association rules on heterogeneous genome data. In *Pacific Symposium on Biocomputing '97*, pages 397–408, 1997.

[21] M. Sauerhoff and I. Wegener. On the complexity of minimizing the OBDD size for incompletely specified functions. *IEEE transactions on computer-aided design of integrated circuits and systems*, 15:1435–1437, 1996.

[22] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proc. of the 21st International Conference on Very Large Data Bases*, pages 432–444, 1995.

[23] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proc. of the 21th International Conference on Very Large Data Bases*, pages 407–419, 1995.

[24] H. Toivonen. Sampling large databases for association rules. In *Proc. of the 22nd International Conference on Very Large Data Bases*, pages 134–145, 1996.

[25] S. Wu and U. Manber. Fast text searching with errors. Technical Report 91-11, Department of computer science, University of Arizona, 1991.