Mining Binary Diagram Rules

Kazuya Inoue [1] Osamu Maruyama [2] Takayoshi Shoudai [1] Satoru Miyano [2]

{kazuya,shoudai}@i.kyushu-u.ac.jp

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

[1] Department of Informatics, Graduate School of Information Science and Electrical Engineering,
Kyushu University, Kasuga 816, Japan

[2] Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

1 Introduction

Mining association rules was introduced as a novel framework of Data Mining by Agrawal *et al.* [1] and some efficient heuristic algorithms for mining all association rules have been devised [2]. Based on such research, some systems for mining association rules have been already in commerce and have succeeded in discovering a rule like *"if a customer buys a swimsuit and a pair of flippers then the customer buys goggles too in 90% of the cases"* from huge databases. This method is obviously applicable to genomic data and a successful result has been reported for finding simple association rules from signals found in mammalian promoter sequences [4].

In order to increase the diversity and expressibility of rules, we introduce a rule in a more general form, *Binary Diagram Rule* (BDR, shortly), so that a rule like *"at teatime, if a person drinks a cup of coffee OR tea, then the person does NOT drinks a glass of milk in 95% of the cases"* is easily expressed.

BDR is defined in terms of binary decision diagrams (BDD), which are decision diagrams such that the internal nodes are labeled with variables and the leaves are labeled with 0 or 1. BDDs represent boolean functions in a compact way and have been originally used for logical circuit design. The flexibility of BDD enables BDR to express more complicated relations and to provide possibilities to discover valuable knowledge from genomic databases.

2 Methods

Some specific patterns are found in a family of amino acid sequences and DNA sequences. Some patterns are known to repeat several times in sequences. For a collection O of sequences and a collection L of patterns, we consider a table $D = (O, L, f)$ whose row and column are indexed by L and O, respectively. The (s, π)-entry is a value $f(s, \pi)$ which represents the number of occurrences of a pattern π in a sequence s. An example is given in Table 1.

A *binary diagram* (BD) is a BDD such that the labels of the leaves are ignored. A BDR consists of two BDs, which are called the *Upper Binary Diagrams* (UBD) and the *Lower Binary Diagrams* (LBD), respectively.

The UBD works as a premise and the LBD stands for a conclusion. Nodes of out-degree 0 are called *terminal nodes*. A BDR is made by linking one of the terminal nodes of UBD with the root node of LBD. In addition, the LBD has a *goal node* as a special terminal node. Each sequence in a table traces a path from the root node to a terminal node according to the answer to a question on each node.

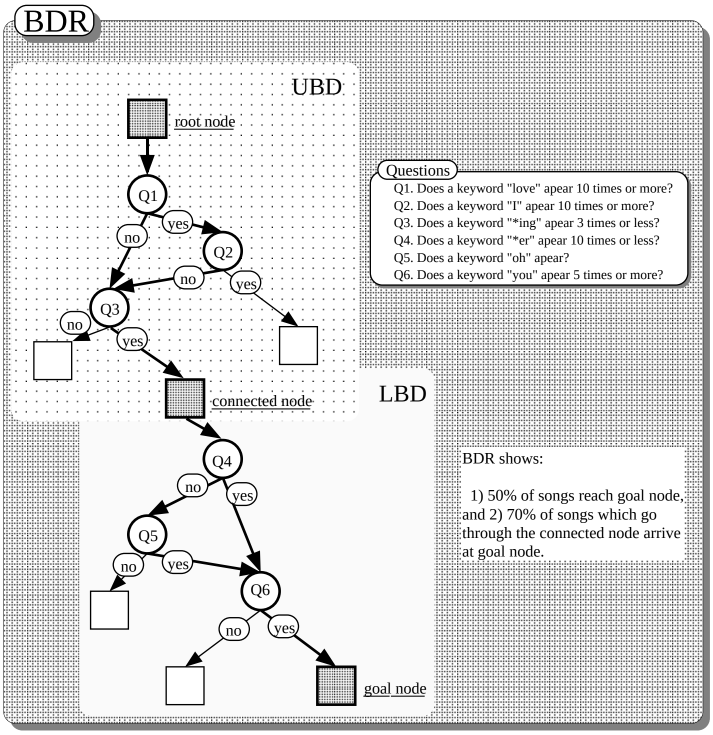
We say that a BDR is good if most of the sequences which go through the connected node of the UBD reach the goal node of the LBD. BDRs visualize the knowledge in the rules. Association rules of Agrawal *et al.* [1] can be also expressed by BDRs.

Our strategy for mining BDRs from a table D is made up of the following two stages: First, an LBD is constructed by a probabilistic method. Next, by using an ID3-like algorithm [5] based on the LBD, a UBD which forms a good BDR is searched. Finally, the method [3] of reducing ordered BDDs is applied to the BDR to simply its structure.

For example, knowledge is represented as a BDR in a way of Fig. 1 for the data of Table 1.

	I	love	you	her	*re	it	shake	oh	*ing	*er
I Saw Her Standing There	12	2	1	13	7	0	0	3	1	15
Misery	2	0	0	6	7	0	0	6	3	19
Anna	19	7	14	15	16	0	0	4	4	20
Chains	5	6	8	0	6	0	0	5	0	0
Boys	5	0	0	2	1	0	0	0	1	4
Ask Me Why	24	6	13	2	8	6	0	0	6	12
Please Please Me	4	2	8	1	3	1	0	0	0	3
Love Me Do	4	25	9	0	0	0	0	1	0	0
P.S. I Love You	13	17	30	2	11	0	0	3	2	10
Baby It's You	6	2	14	0	0	8	0	3	0	7
Do You Want to Know a Secret	2	4	18	0	7	0	0	6	0	8
A Taste of Honey	2	0	4	1	7	0	0	0	4	5
There's A Place	8	1	6	9	9	4	0	0	2	9
Twist and Shout	2	0	20	0	4	26	12	0	0	4

Table 1: The Beatles, "Please Please Me".



Q1. Does a keyword "love" apear 10 times or more?
Q2. Does a keyword "I" apear 10 times or more?
Q3. Does a keyword "*ing" apear 3 times or less?
Q4. Does a keyword "*er" apear 10 times or less?
Q5. Does a keyword "oh" apear?
Q6. Does a keyword "you" apear 5 times or more?

BDR shows:

1) 50% of songs reach goal node, and 2) 70% of songs which go through the connected node arrive at goal node.

Figure 1: Good BDR which is mined from Table 1.

Acknowledgment

The authors would like to thank A. Shinohara, S. Matsumoto and N. Furukawa for useful comments.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", *ACM SIGMOD*, pp. 207–216, 1993.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Fast Algorithms for Mining Association Rules", *IBM Technical Report RJ9839*, 1994.

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

[4] G. Shibayama, K. Satou, and T. Takagi, "Mining Association Rules from Signals found in Mammalian Promoter Sequences", *Proceedings of Genome Informatics Workshop*, pp. 108–109, 1995.

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