BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 6(5)

A heuristic method for discovering biomarker candidates based on rough set theory

Yasuo Kudo* & Yoshifumi Okada

College of Information and Systems, Muroran Institute of Technology, 27-1 Mizumoto, Muroran, Hokkaido 050-8585, Japan; Yasuo Kudo - Email: kudo@csse.muroran-it.ac.jp; Phone: +81 143 46 5469; Fax: +81 143 46 5499; *Corresponding author

Received May 13, 2011; Accepted May 23, 2011; Published May 26, 2011

Abstract:
We apply a combined method of heuristic attribute reduction and evaluation of relative reducts in rough set theory to gene expression data analysis. Our method extracts as many relative reducts as possible from the gene-expression data and selects the best relative reduct from the viewpoint of constructing useful decision rules. Using a breast cancer dataset and a leukemia dataset, we evaluated the classification accuracy for the test samples and biological meanings of the rules. As a result, our method presented superior classification accuracy comparable to existing salient classifiers. Moreover, our method extracted interesting rules including a novel biomarker gene identified in recent studies. These results indicate the possibility that our method can serve as a useful tool for gene expression data analysis.

Background:

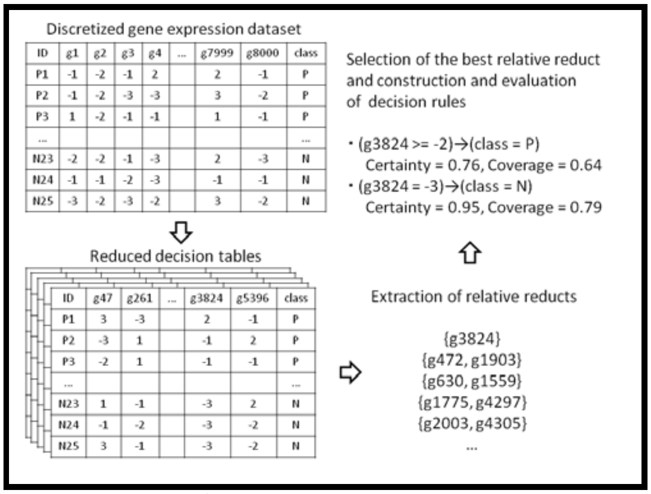
DNA microarray technology has enabled us to monitor the expression levels of thousands of genes simultaneously under certain conditions, and has been yielded various applications in the field of disease diagnosis **[1]**, drug discovery **[2]**, and toxicological research **[3]**. Among them, cancer informatics based on gene-expression data is an important domain that has promising prospects for both clinical treatment and biomedical research. One of the key issues in this domain is to discover biomarker genes for cancer diagnosis from a massive amount of gene-expression data by using a bioinformatics approach called gene selection. A typical gene-selection approach is a statistical test such as t-test and ANOVA **[4]**. This approach detects differentially expressed genes between groups of samples obtained from different cells/tissues. Most of the statistical tests assume that the expression values of each gene across the samples follow a prior probability distribution; hence a sufficiently large number of samples are required to obtain statistically reliable results. Rough set theory **[5]** provides a theoretical basis for set-theoretical approximation and rule generation from categorical data. Computation of relative reducts is one of the hottest and most important research topics in rough set theory as a basis for rule generation. Relative reducts are minimal sets of attributes for correctly classifying all samples to those classes. We then expect that computation of relative reducts from gene-expression data is useful for discovering biologically-meaningful information such as biomarker candidates for cancer diagnosis. Because computing all relative reducts of the given data requires very high computational cost, there have been many proposals of heuristic algorithms to compute some of the candidates of relative reducts **[6-10]**. Kudo and Murai proposed attribute-reduction algorithms to compute as many relative reducts as possible from a decision table with numerous condition attributes **[11]**. They also proposed an evaluation criterion of relative reducts that evaluates the usefulness of relative reducts from the viewpoint of decision-rule generation **[12]**. In this paper, we introduce Kudo and Murai's heuristic attribute reduction algorithms **[11]** and a criterion of relative reducts **[12]** for gene-expression data analysis. We use these algorithms and criterion in two gene-expression datasets, breast cancer **[13]** and leukemia **[14]**, and discuss the extracted decision rules from these datasets and their biological meanings. The experimental results indicate that the method used in this paper can identify differentially expressed genes between different classes in gene-expression datasets and that it can be useful for gene-expression data analysis.

Methodology:

The method we use in this paper to extract decision rules from gene-expression data based on rough set theory consists of the following three components: (1) Extraction of as many relative reducts as possible from gene-expression data; (2) Selection of relative reducts in accordance with an evaluation criterion of relative reducts; (3) Construction of decision rules from the selected relative reducts. **Figure 1** illustrates the processing flow of our method. In the following section, in terms of the method we use in this paper, we introduce heuristic attribute-reduction algorithms for generating as many relative reducts as possible **[11]** used in the first step of the above method and a criterion for evaluating the usefulness of relative reducts **[12]** as in the second step. Note that the details of these algorithms and the criterion of relative reducts are in **Supplementary material**.

Datasets:

To evaluate the usefulness of our method, we use two gene-expression datasets: breast cancer **[13]** and leukemia **[14]**. Both of them are two-class datasets. The leukemia dataset is composed of the gene-expression values for 12,582 genes in 24 acute lymphocytic leukemia (ALL) samples and 28 acute myeloid leukemia (AML) samples. The breast cancer dataset includes the gene-expression values for 7,129 genes in 25 positive and 24 negative samples. For each dataset, the expression values from each gene are linearly normalized to have mean 0 and variance 1. Subsequently, they are discretized into six bins (-3, -2, -1, 1, 2, 3) by uniformly dividing the difference between the maximum and the minimum in the normalized data and into one bin that represents the lack of gene-expression values. Discretized positive values represent that the genes are up-regulated, while negative values represent that genes are down-regulated.



Discretized gene expression dataset

ID	g1	g2	g3	g4	...	g7999	g8000	class
P1	-1	-2	-1	2		2	-1	P
P2	-1	-2	-3	-3		3	-2	P
P3	1	-2	-1	-1		1	-1	P
...								...
N23	-2	-2	-1	-3		2	-3	N
N24	-1	-1	-2	-3		-1	-1	N
N25	-3	-2	-3	-2		3	-2	N

Reduced decision tables

ID	g47	g261	...	g3824	g5396	class
P1	3	-3		2	-1	P
P2	-3	1		-1	2	P
P3	-2	1		-1	-1	P
...						...
N23	1	-1		-3	2	N
N24	-1	-2		-3	-2	N
N25	3	-1		-3	-2	N

Extraction of relative reducts

{g3824}
{g472, g1903}
{g630, g1559}
{g1775, g4297}
{g2003, g4305}
...

Selection of the best relative reduct and construction and evaluation of decision rules

- (g3824 >= -2)→(class = P)
  Certainty = 0.76, Coverage = 0.64
- (g3824 = -3)→(class = N)
  Certainty = 0.95, Coverage = 0.79

Figure 1: The method of discovering biomarker candidates based on rough set theory

Results and Discussion:

Parameters:

Our method was implemented in Java on a Linux workstation (CPU: Intel Xeon X5460 (3.16GB) x2, Memory: 8GB, HDD: 160GB, OS: SUSE Linux 10.1). All experiments were conducted with the following parameters: base size $b = 10$, size limitation $L = 25$, and number of iterations $I = 100$.

Classification Accuracy:

First, we evaluate the classification accuracy of our method. The evaluation is conducted by Leave-One-Out Cross Validation (LOOCV). In LOOCV, first, we extract one sample as a test sample from the dataset and generate rules using the remaining samples. Second, we check whether the test sample is correctly classified by the rules. These processes are repeated for all samples. Finally we calculate the rate of correctly classified samples. The classification accuracy is compared to those of the two salient classifiers, decision tree (C4.5) and support vector machine (SVM). **Table 1 (see Supplementary material)** shows the results of LOOCV on our method, C4.5, and SVM. For the breast cancer dataset, our method exhibits the similar classification ability with C4.5 and SVM. For the leukemia dataset, the classification ability of our method exceeds greatly that of C4.5.

Biological meanings of extracted rules:

Next, we discuss the biological meanings of the best results by applying our method 10 times for each dataset. In these experiments, we used the same parameter settings with the comparison experiments. The best relative reducts of two datasets are as follows: (1) Breast cancer dataset: {*CRIP1*, *M34715_at*}, *ACov* = 0.08 (= 2/26). (2) Leukemia dataset: {*POU2AF1*}, *ACov* = 0.29 (= 2/7), where the score *ACov* is the average of coverage of decision rules generated from the relative reduct defined by Eq.(2) in **Supplementary material**. For example, the relative reduct {*POU2AF1*} of leukemia dataset generates 7 decision rules from 2 classes, i.e., AML and ALL; hence *ACov* score of the relative reduct {*POU2AF1*} is 2/7 (= 0.29). We extracted rules from each dataset by performing the following three steps: 1) generating all decision rules by the best relative reduct of each dataset, 2) removing decision rules that contain null values in the antecedents, and 3) combining the generated decision rules as long as possible by interpreting the meanings of decision rules. As a result, we obtained the rules for each dataset. **(see Supplementary material)**

The extracted rules are evaluated on the basis of known biological findings. To this end, we investigate the functions of genes in the rules by reference to a genetic disease database (OMIM) **[15]** and a protein sequence database (Swiss-Prot) **[16]**. For the breast-cancer dataset, the samples can be discriminated into a true class with an accuracy of 88 percent according to the expression level of the Cystein-rich intestinal protein 1 (*CRIP1*). *CRIP1* is a transcription-factor gene that induces apoptosis in cancer cells. Interestingly, this gene has been identified as a novel biomarker of human breast cancer in recent studies **[17, 18]**. In the extracted rule, we can see that the *CRIP1* expression is more up-regulated in the positive samples. Indeed, this is consistent with the recent findings by Ma *et al.* **[17]** that *CRIP1* in human breast cancer was over-expressed, compared to normal breast tissue in *in situ* experiments. For the leukemia dataset, all samples can be perfectly discriminated by the expression level of the POU class 2 associating factor 1 (*POU2AF1*). *POU2AF1* is known as a gene responsible for leukocyte differentiation. In Swiss-Prot, we can see the description that "a chromosomal aberration involving *POU2AF* may be a cause of a form of B-cell leukemia." Namely, it suggests that this gene can be inactivated/down-regulated in lymphocytic leukemia, such as ALL. In contrast, it should be noted that *POU2AF1* in the extracted rule shows a weaker expression in AML than ALL. At present, the detailed role of *POU2AF1* in AML has not been revealed **[19]**, whereas we expect that its biological relevance will be unveiled by experimental biologists in the near future.

Conclusion:

In this paper, we introduced a combined method of heuristic attribute reduction and evaluation of relative reducts in rough set theory for gene-expression data analysis. Our method is based on a heuristic attribute-reduction algorithm for generating as many relative reducts as possible and a criterion for evaluating the usefulness of relative reducts. We applied our method to two gene-expression datasets: breast cancer and leukemia. In the comparison of our method with C4.5 and SVM, our method showed good classification accuracy that is comparable to the results of SVM and considerably exceeds that of C4.5.The experimental results also showed that our method can identify differentially expressed genes among different classes in gene-expression datasets. For the breast-cancer dataset, our method extracted decision rules regarding a gene that has been identified as a novel biomarker of human breast cancer in recent studies. For the leukemia dataset, rules about a gene responsible for leukocyte differentiation were extracted. Thus, these results indicate a possibility that our method can be a useful tool for gene-expression data analysis. As a future work, we reduce the computation time via improvement of the program and will provide the freely available tool.

References:

[1] Yoo SM *et al. J Microbiol Biotechnol*. 2009 **19**: 635 [PMID: 19652509]

[2] Debouck C & Goodfellow PN. *Nat Genet*. 1999 **21**: 48 [PMID: 9915501]

[3] Vrana KE *et al. Neurotoxicology* 2003 **24**(3): 321 [PMID: 12782098]

[4] Cui X & Churchill GA. *Genome Biol*. 2003 **4**(4): 210 [PMID: 12702200]

[5] Pawlak Z. *Int J Computer and Information Science*. 1982 **11**: 341

[6] Guan JW & Bell DA. *Artificial Intelligence*. 1998 **105**: 77

[7] Chouchoulas A & Shen A. *Applied Artificial Intelligence*. 2001 **15**(9): 843

[8] Kryszkiewicz M & Lasek P. *Transactions on Rough Sets*. 2008 **9**: 76

[9] Yao YY *et al. Transactions on Rough Sets*. 2008 **8**: 332

[10] Kudo Y & Murai T. *J Advanced Computational Intelligence and Intelligent Informatics*. 2011 **15**(1): 102

[11] Kudo Y & Murai T. Proc. IEEE GrC2010. 2010; 265-270.

[12] Kudo Y & Murai T. *Int J Cognitive Informatics and Natural Intelligence*. 2010 **4**(2): 50

[13] West M *et al. Proc Natl Acad Sci U S A*. 2008 **98**(20): 11462 [PMID: 11562467]

[14] Armstrong SA *et al. Nat Genet*. 2002 **30**(1): 41 [PMID: 11731795]

[15] http://www.nslij-genetics.org/search_omim.html

[16] http://au.expasy.org/sprot/

[17] Ma XJ *et al. Proc Natl Acad Sci U S A*. 2003 **100**(10): 5974 [PMID: 12714683]

[18] Liu S *et al. Mol Cancer Res*. 2004 **2**(8): 477 [PMID: 15328374]

[19] Gibson SE *et al. Am J Clin Pathol*. 2006 **126**(6): 916 [PMID: 17074681]

Edited by P Kangueane
Citation: Kudo & Okada. Bioinformation 6(5): 200-203 (2011)

Supplementary material:

Table 1: Comparison of classification accuracy of the proposed method, decision tree, and SVM

	Classification accuracy (%)	
	Breast cancer	Leukemia
Proposed method	82.86	92.86
Decision tree (C4.5)	83.67	73.61
SVM (Linear kernel)	87.75	97.22

Breast cancer dataset:

$(CRIP1 \geq -2) \rightarrow (\text{Class} = \text{Positive})$, Certainty = 0.76, Coverage = 0.64.
$(CRIP1 = -3) \rightarrow (\text{Class} = \text{Negative})$, Certainty = 0.95, Coverage = 0.79.

Note that the gene *M34715_at* was removed in combining generated decision rules from the best relative reduct {*CRIP1*, *M34715_at*} because, in this case, *M34715_at* was used only for classifying one positive subject with *CRIP1* = -3.

Leukemia dataset:

$(POU2AF1 \geq -2) \rightarrow (\text{Class} = \text{ALL})$, Certainty = 1.0, Coverage = 0.88.
$(POU2AF1 = -3) \rightarrow (\text{Class} = \text{AML})$, Certainty = 1.0, Coverage = 1.0.

Heuristic Algorithms for Attribute Reduction Using Reduced Decision Tables:

We review heuristic algorithms to generate as many relative reducts as possible from decision tables with numerous condition attributes proposed by Kudo and Murai **[11]**. These heuristic algorithms are based on the idea of reduced decision tables that preserve the discernibility of objects that belong to mutually different decision classes in the given decision table. Formally, a reduced decision table of a given decision table is defined as follows.

Definition 1. Let $DT = (U, C, d)$ be a decision table, where U is a finite and non-empty set of objects, C is a finite and non-empty set of condition attributes, and d is a decision attribute. A reduced decision table of DT is the following triple:

$$RDT = (U, C', d) \qquad (1)$$

where U and d are identical to DT. The set of condition attributes C'satisfies the following conditions:

1. $C' \subseteq C$.
2. For any objects x_i and x_j that belong to different decision classes, if x_i and x_j are discernible by the indiscernibility relation R_C based on C, the two objects x_i and x_j are also discernible by the indiscernibility relation $R_{C'}$ based on C'.

Algorithm 1 below generates a reduced decision table of the given decision table. In Algorithm 1, the condition attributes are selected from C at random based on the parameter of base size b that decides the minimum number of condition attributes of the reduced decision table; some attributes in the elements of the discernibility matrix are supplied to preserve the discernibility of the objects in the given decision table.

Algorithm 1 Decision-table reduction algorithm

```
Input: decision table $DT = (U, C, d)$, discernibility matrix $DM$ of $DT$, base size $b$
Output: reduced decision table $RDT = (U, C', d)$
Select $b$ attributes $a_1, \cdots, a_b$ from $C$ at random by sampling without replacement.
$C' = \{a_1, \cdots, a_b\}$
For all $\delta_{ij} \in DM$ such that $i > j$ do
    If $\delta_{ij} \neq \emptyset$ and $\delta_{ij} \cap C' = \emptyset$ then
        Select $c \in \delta_{ij}$ at random
        $C' := C' \cup \{c\}$
    End if
End if
Return $RDT = (U, C', d)$
```

Note that, for any decision table and any reduced decision table, a set of condition attributes A is a relative reduct of the reduced decision table if and only if A is also a relative reduct of the given decision table. Thus, generating as many reduced decision tables and relative reducts as possible from the given decision table, we can extract many relative reducts from the given decision table. Algorithm 2 below generates the relative reducts of the given decision table based on generating reduced decision tables by Algorithm 1 and switching the exhaustive attribute reduction and heuristic attribute reduction according to the number of condition attributes of each reduced decision table.

Algorithm 2 Exhaustive / heuristic attribute-reduction algorithm

```
Input: decision table $DT = (U, C, d)$, base size $b$, size limit $L$, number of iteration $I$
```

```
Output: set of relative reduct candidates $RED$
$RED = \varnothing$
Construct $DM$ of $DT$
If $|C| \leq L$ then
    $RED =$ Set of all relative reducts by exhaustive attribute reduction from $DT$
Else
    For $i = 1$ to $I$
        $RDT =$ Reduced decision table by Algorithm 1 with $DT$, $DM$, and $b$
        If $|C'| \leq L$ then
            $S =$ Set of all relative reducts by exhaustive attribute reduction from $RDT$
        Else
            $S =$ Set of relative reduct candidates by heuristic attribute reduction from $RDT$
        End if
        $RED := RED \cup S$
    End for
End if
Return $RED$
```

In Algorithm 2, the size limit L is the threshold for switching attribute-reduction methods, and if the number of condition attributes of a decision table is smaller than L, Algorithm 2 tries to generate the set of all relative reducts of the decision table. Thus, we need to set the threshold L appropriately. If the number of condition attributes of the given decision table DT is greater than the threshold L, Algorithm 2 repeats I times generating the reduced decision table RDT and the attribute reduction from RDT by selecting the exhaustive method or the heuristic method, and generates the set RED of relative reducts. Note that RED may contain some outputs with redundancy if the result of the heuristic attribute reduction is not guaranteed to generate relative reducts.

An Evaluation Criterion of Relative Reducts:

From the viewpoint of data analysis using rough set theory, Kudo and Murai **[12]** proposed an evaluation criterion of relative reducts for extracting useful decision rules. This criterion evaluates the usefulness of each relative reduct by the average of the coverage of the decision rules generated from the relative reduct. Let $DT = (U, C, d)$ be a decision table. For any non-empty set $B \subseteq C$ of condition attributes, the average of coverage $ACov(B)$ of all decision rules generated from B is calculated as follows **[12]**:

$$ACov(B) = \frac{|\mathbf{D}|}{\sum_{[x]_B \in U/R_B} \left| \left\{ D_j \in \mathbf{D} \,\middle|\, [x]_B \cap D_j \neq \varnothing \right\} \right|}, \qquad (2)$$

where $[x]_B$ is the equivalence class of element $x \in U$ by the equivalence relation R_B, U / R_B is the quotient set of U by R_B, D_j is a decision class, $\mathbf{D}$ is the set of all decision classes, and $|X|$ is the cardinality of set X. For any non-empty set $B \subseteq C$ of condition attributes, the range of score $ACov(B)$ is $0 < ACov(B) \leq 1$ and $ACov(B) = 1$ holds if each decision class is completely described by only one decision rule, or equivalently, the classification result using B is identical to the classification result using the decision attribute d.

Equation (2) indicates that $ACov(B)$ is the number of decision classes over the number of decision rules generated from B and depends only on the number of decision rules, because the number of decision classes is fixed in the given decision table. For each relative reduct $E \subseteq C$ of the given decision table, the average of coverage $ACov(E)$ reflects the roughness of partition U / R_E by the equivalent classes based on R_E. It is guaranteed that, for any relative reducts $E, F \subseteq C$, if the partition U / R_E is rougher than the partition U / R_F, then $ACov(E) \geq ACov(F)$ holds and this property provides a theoretical basis for using Eq.(2) as an evaluation criterion of relative reducts.