An Evolutionary Associative Contrast Rule Mining Method for Incomplete Database

Kaoru Shimada and Takashi Hanioka
Fukuoka Dental College, 2-15-1 Tamura, Sawara, Fukuoka, 814-0193, Japan

Abstract—*A method for associative contrast rule mining from incomplete database is demonstrated to find interesting differences between two incomplete data sets. The method extracts rules like "if X then Y" is interesting only in the focusing class. The method has been developed using a basic structure of the evolutionary graph-based optimization technique and adopting a new evolutionary strategy to accumulate rules through its evolutionary process. The method can realize the association analysis between two classes of the incomplete database using chi-square test. We evaluated the performance of the evolutionary method for associative contrast rule mining for the incomplete database. In addition, the evaluation of the mischief for the rule measurements by missing values is demonstrated.*

Keywords: association rules, missing values, evolutionary computation and genetic algorithms, contrast mining

1. Introduction

Association rule mining is the discovery of association relationships or correlations among a set of attributes (items) in a database. Association rule in the form of 'if X then Y ($X \rightarrow Y$)' is interpreted as 'the set of attributes X are likely to satisfy the set of attributes Y'. Many techniques for association rule mining and its applications have been proposed, which achieve quite effective performances [1], [2]. However, previous approaches cannot handle incomplete database. An incomplete database includes missing values in some instances. For example, the database of questionnaires probably includes missing data such as age, income, and so on. In the case plural databases are joined, missing data would also appear because attributes in each database are not the same. Conventional rule mining methods regard the database as complete, or disregard instances including missing values. Instances including missing data are deleted for rule mining or filled in with the mean values or frequent category [3], [4]. When the data sets have a huge number of instances, it is easy to take these policies. However, the data mining for dense database like medical data is different from the situation. Experimental data sets probably include missing values caused by the failure of the experiments or extraordinary values. It is not possible for these cases to fill the missing values with mean values or frequent categories.

We have already proposed an association rule mining method for incomplete database using an evolutionary computation technique [5], [6]. The method extract rules directly without constructing the frequent itemsets used in the previous approaches. Available attribute values in an instance including missing values are used for the calculation of rule measurements. The method has been developed using a basic structure of Genetic Network Programming (GNP) and adopting a new evolutionary strategy to accumulate rules through its evolutionary process. GNP is one of the evolutionary optimization techniques, which uses the directed graph structures as genes [7], [8]. Conventional Genetic Algorithm (GA) based methods extract a small number of rules optimizing a given fitness function [9], [10]. On the other hand, in the GNP based method, rules satisfying given conditions are accumulated in a rule pool through GNP generations and extracted rules are reflected in genetic operators as acquired information. GNP individuals evolve in order to store new interesting rules into the pool as many as possible, not to obtain the individual with highest fitness.

In this paper, the GNP based rule mining method is extended to the associative contrast rule mining to find interesting differences between two incomplete data sets. The associative contrast rule is defined as follows: although $X \rightarrow Y$ satisfies the given importance conditions within Database A, however, the same rule $X \rightarrow Y$ does not satisfy the same conditions within Database B [7]. The method can realize the association analysis between two classes of the incomplete database using χ^2 test. When we use the conventional rule extraction methods, it is not easy to extract such rules, because we have to check the combinations of rules one by one. In [5], the algorithm for rule mining from incomplete database was proposed, however, such as the comparison of the performance of the rule extraction and the mischief for the rule measurements by missing values were not evaluated sufficiently. In this paper, we describe the performance of the evolutionary method of associative contrast rule mining for incomplete database. In addition, the evaluation of the mischief for the rule measurements by missing values is demonstrated.

This paper is organized as follows: in the next section, some related concepts on associative contrast rules in the incomplete database are presented. In Section 3, an algorithm capable of finding the associative contrast rules from the incomplete database is described. Experimental results are presented in Section 4, and conclusions are given in Section 5.

2. Associative Contrast Rules

Let A_i be an attribute in the database and C be the class labels. The attribute values of tuples are indicated by 1 or 0 as shown in Table 1 (a). The absence of item A_i is described as $A_i = 0$ and missing data (lack of information) are indicated as 'm' different value from 1 and 0. For example, $ID = 4$ in Table 1 (a) misses the data of attribute A_2. In this paper, we use database form like Table 1 (a). Suppose that the class label is $C = 1$ or $C = 0$, that is, the database is divided into two classes, and the database has no missing data in the class label.

X and Y denote the following combinations of attributes: $X = (A_j = 1) \wedge \cdots \wedge (A_k = 1), Y = (A_m = 1) \wedge \cdots \wedge (A_n = 1)$, $X \cap Y = \emptyset$. X is represented briefly as $A_j \wedge \cdots \wedge A_k$. An association rule is an implication of the form $X \rightarrow Y$. X is called antecedent and Y is called consequent of the rule. If the number of tuples containing X in the database equals x, then we define $\alpha = support(X) = x/N$, where, N is the total number of tuples for the rule evaluation. Let $\beta = support(Y) = y/N$ and $\gamma = support(X \wedge Y) = z/N$ using y and z, the number of tuples containing Y and $X \wedge Y$, respectively. The rule has measures of its frequency called $support$ and its strength called $confidence$ defined by

$$support(X \rightarrow Y) = \frac{z}{N}, confidence(X \rightarrow Y) = \frac{z}{x}.$$

In addition, the significance of association via the chi-square test for correlation used in classical statistics is also used for the measurement. χ^2 value of the rule $X \rightarrow Y$ is given as

$$\chi^2(X \rightarrow Y) = \frac{N(\gamma - \alpha\beta)^2}{\alpha\beta(1-\alpha)(1-\beta)}. \qquad (1)$$

In the case of the rule extraction from incomplete database, the number of tuples for measurement calculation is different rule by rule [5]. For example, let $(A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1) \rightarrow (A_4 = 1)$ be a candidate rule in Table 1. It is clear that the tuple $ID = 1$ in Table 1 does not satisfy this rule by $A_2 = 0$. When at least one of the values of A_1, A_2, A_3 or A_4 equal 0, it is sure that the tuple does not satisfy the rule. Therefore, $ID = 4, 5$ and 6 are available to judge for the rule even if they have missing values. These tuples are available for the calculation of rule measurements. However, tuples $ID = 7$ and 8 are not available, because we cannot judge whether the tuples satisfy the rule or not by missing values. Therefore, the tuples whose attribute values equal 1 or m, but not the tuples whose all attribute values equal 1 should be excluded. Measurements of the above rule are

$$support((A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1) \rightarrow (A_4 = 1)) = \frac{1}{6},$$

$$confidence(A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1) \rightarrow (A_4 = 1)) = \frac{1}{1}.$$

Table 1: An example of incomplete database.

(a)

ID	A_1	A_2	A_3	A_4	C
1	1	0	1	0	0
2	1	1	1	1	1
3	1	1	0	0	1
4	0	m	1	1	0
5	0	m	m	1	0
6	m	m	1	0	1
7	1	1	1	m	1
8	m	1	1	m	1

(b)

$A_1 \wedge A_2 \wedge A_3 \rightarrow A_4$
not satisfy
satisfy
not satisfy
not satisfy
not satisfy
not satisfy
cannot judge
cannot judge

In this paper, missing rate is defined as the ratio of the number of missing values and the total number of attribute values. In Table 1, for example, 8 missing values are found within 32 values of A_1, A_2, A_3 and A_4, then, missing rate is 8/32=0.25 (25%). M value and Y value introduced in [5] are used for the measurements calculation of rules as follows. M value represents the number of tuples whose attribute values for the rule are equal 1 or m, and Y value represents the number of tuples whose attribute values for the rule are all equal to 1. N value which is the number of available tuples is also defined for the rule measurement calculation. For example, M value for the above rule equals 3 ($ID = 2, 7$ and 8). Y value is 1 ($ID = 2$) and N value is 6. These values satisfy the following formula:

$$N \text{ value} = N_T - (M \text{ value} - Y \text{ value}),$$

where, N_T is the total number of tuples in the database. When the database is complete, N value equals N_T.

In the case of data mining from the dense database, such as the medical data, differences between two data sets gathered by different conditions are more interesting than support-confidence framework. The following rule showing difference between class labels [7] is considered.

[Associative contrast rule] Although $X \rightarrow Y$ satisfies the given importance conditions within $C = 1$, $X \rightarrow Y$ does not satisfy the conditions within $C = 0$.

For example, conditions of importance for associative contrast rules are defined using chi-square value as follows:

$$\chi^2(X \rightarrow Y)_{(C=1)} > \chi^2_{min} \qquad (2)$$

$$\chi^2(X \rightarrow Y)_{(C=0)} < \chi^2_{max} \qquad (3)$$

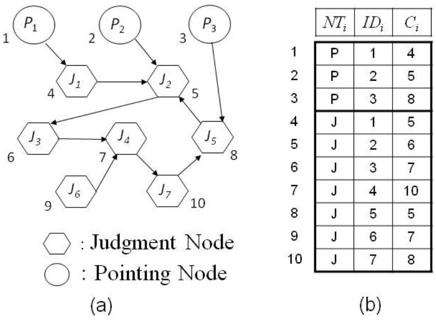
$$support(X \rightarrow Y)_{(C=1)} \geq supp_{min}, \qquad (4)$$

$$support(X \rightarrow Y)_{(C=0)} \geq supp_{min}, \qquad (5)$$

where, χ^2_{min} and χ^2_{max} ($\chi^2_{min} \geq \chi^2_{max}$) and $supp_{min}$ are the threshold values given by users in advance. $C = 1$ and $C = 0$ represent the focused class label for the rule. It is not easy for the conventional frequent itemset based methods to extract the above rules, because we have to check the combinations of rule measurements one by one.

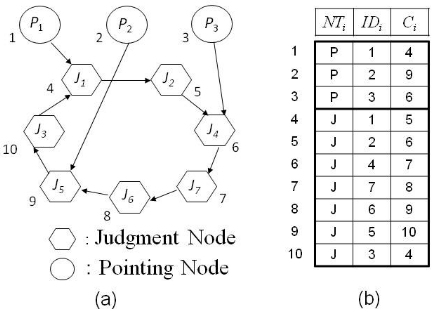
Instead of (2) and (3), like the following conditions on the threshold for $confidence$ could be used.

$$confidence(X \rightarrow Y)_{(C=1)} - confidence(X \rightarrow Y)_{(C=0)} > \delta \qquad (6)$$



	NT_i	ID_i	C_i
1	P	1	4
2	P	2	5
3	P	3	8
4	J	1	5
5	J	2	6
6	J	3	7
7	J	4	10
8	J	5	5
9	J	6	7
10	J	7	8

Fig. 1: Basic structure of individual in GNP-based method.



	NT_i	ID_i	C_i
1	P	1	4
2	P	2	9
3	P	3	6
4	J	1	5
5	J	2	6
6	J	4	7
7	J	7	8
8	J	6	9
9	J	5	10
10	J	3	4

Fig. 2: Basic structure of individual in ring structure method.

$$confidence(X \rightarrow Y)_{(C=0)} \\ - confidence(X \rightarrow Y)_{(C=1)} > \delta \quad (7)$$

where, δ is a constant expressing the threshold of the difference of $confidence$.

3. Evolutionary Rule Mining Method

In this section, the associative contrast rule mining method for incomplete database based on evolutionary computation is described [5]. The form of rules and conditions of threshold values for interestingness are given by users in advance. Rule representations and fitness function are designed based on the users objects. The task for rule extraction is done accumulatively through evolutionary process, not to obtain elite individual at the final generation.

3.1 Structure of Individuals

The basic structure of the GNP individual is shown in Fig. 1. the individual is composed of two kinds of nodes: Judgment node and Pointing node (Processing nodes in [5] are renamed as *Pointing nodes*). P_1 is a Pointing node and is a starting point of rules. Each Pointing node has an inherent numeric order (P_1, P_2, . . . , P_s) and is connected to a Judgment node. Each Judgment node has two connections: Continue-side and Skip-side. The Continue-side of the node is connected to another Judgment node. The Skip-side of the node is connected to the next numbered Pointing node. The Skip-side of Judgment nodes are abbreviated in Fig. 1 (a).

The gene structure of the GNP individual is shown in Fig. 1 (b). NT_i describes the node type and ID_i is an identification number of functions. C_i denotes the nodes ID which are connected from node i as Continue-side. All individuals in a population have the same number of nodes.

In this paper, *Ring structure* method and *Random network* method are introduced for the purpose of the comparison. *Ring structure* utilizes an individual using the same settings as GNP except the Judgment node connection is restricted to make ring structure, that is, one ring form is composed using all the Judgment nodes (See Fig. 2). *Random network* utilizes an individual using the same settings of GNP except the evolutionary mechanism. The connections and functions of Judgment nodes are initialized every generation.

3.2 Basic Idea of Rule Representation

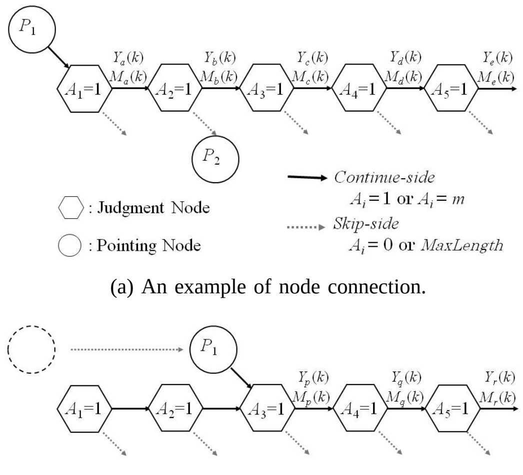
Rules are represented as the connections of nodes in an individual. Attributes and their values correspond to the functions of Judgment nodes. Fig. 3 (a) shows a sample of the node connection in the individual. '$A_1 = 1$', '$A_2 = 1$', '$A_3 = 1$', '$A_4 = 1$' and '$A_5 = 1$' in Fig. 3 (a) denote the functions of Judgment nodes. The connections of these nodes represent rules like $(A_1 = 1) \rightarrow (A_2 = 1)$ and $(A_1=1) \wedge (A_2=1) \rightarrow (A_3=1)$.

Judgment nodes can be reused and shared with some other rule representations because of the GNP's feature. GNP individual generates many rule candidates using its graph structure. The kinds of the Judgment node functions equal the number of attributes in the database.

If a rule symbolized by node connections is interesting, then the rules symbolized by after changing the connections or functions of nodes could be candidates of interesting ones. We can obtain these rule candidates effectively by genetic operations for individuals, because mutation or crossover operation change the connections or contents of the nodes.

3.3 Node Transition in the Individual

Individuals examines the attribute values of each tuple using Judgment nodes and calculates the measurements of rules using Pointing nodes. Judgment node determines the next node by a judgment result. When the attribute value equals 1, then we move to the Continue-side. On the other hand, in the case that the attribute value equals 0, the Skip-side is used for the transition. For example, in Table 1 (a), the tuple $1 \in ID$ satisfies $A_1=1$ and $A_2=0$, therefore, the node transition from P_1 to P_2 occurs in Fig. 3 (a). When the attribute value is missing, then, move to the Continue-side. If the transition to Continue-side connection continues and the number of the Judgment nodes from the Pointing node becomes a cutoff value (given maximum number of attributes in rules, $MaxLength$), then, the connection is transferred to the next Pointing node using the Skip-side



(a) An example of node connection.

(b) Change of Pointing node connection.

Fig. 3: An example of node connection for rule mining.

obligatorily. Skip-side of the Judgment node is connected to the next numbered Pointing node. Then, another examinations of attribute values start at the next Pointing node. If the examination of attribute values from the starting point P_s ends, then the individual examines the tuple $2 \in ID$ from P_1 likewise. Thus, all tuples in the database are examined.

3.4 Calculation of Rule Measurements

Y value and M value are obtained as the numbers of tuples moved to the Continue-side at each Judgment node. These values are counted up and stored in memories. In addition, each Judgment node examines the case of $C = k(k = 0, 1)$ at the same time. In Fig. 3 (a), $Y_a(k)$, $Y_b(k)$, $Y_c(k)$, $Y_d(k)$ and $Y_e(k)$ are the numbers of tuples which belong to class $C = k$ and move to the Continue-side at each Judgment node satisfying that all the attribute values are equal to 1 from the pointing node (Y value). On the other hand, $M_a(k)$, $M_b(k)$, $M_c(k)$, $M_d(k)$ and $M_e(k)$ are the number of tuples at each Judgment node satisfying that the attribute values are equal to 1 or missing values (M value). Using these values, N values, that is, the number of available tuples for the rule measurements calculation are calculated as follows:

$$N_x(k) = N_T - (M_x(k) - Y_x(k)) \tag{8}$$

where, x is a position of the Judgment node. For example, $N_d(k)$ is obtained as $N_d(k) = N_T - (M_d(k) - Y_d(k))$.

Rule measurements are calculated using the above numbers. For example, in the case of $Rule : (A_1 = 1) \wedge (A_2 = 1) \rightarrow (A_3 = 1) \wedge (A_4 = 1)$, the measurements for $C = k$ are

$$support(Rule_{(C=k)}) = \frac{Y_d(k)}{N_d(k)},$$

$$confidence(Rule_{(C=k)}) = \frac{Y_d(k)}{Y_b(k) - (N_b(k) - N_d(k))}.$$

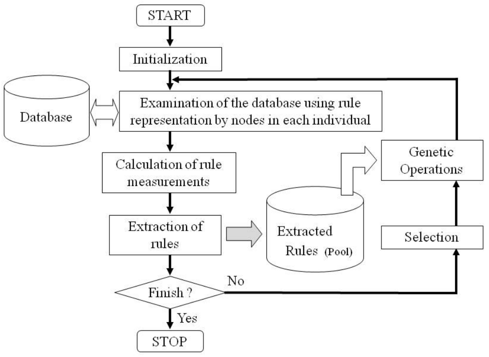

START
Initialization
Database
Examination of the database using rule representation by nodes in each individual
Calculation of rule measurements
Genetic Operations
Extraction of rules
Extracted Rules (Pool)
Selection
Finish ?
No
Yes
STOP

Fig. 4: Flow of the GNP-based rule extraction.

Table 2: Measurements of rules within $C = k$ ($k = 0, 1$).

Association Rules	Support	Confidence
$A_1 \rightarrow A_2$	$\frac{Y_b(k)}{N_b(k)}$	$\frac{Y_b(k)}{Y_a(k)-(N_a(k)-N_b(k))}$
$A_1 \rightarrow A_2 \wedge A_3$	$\frac{Y_c(k)}{N_c(k)}$	$\frac{Y_c(k)}{Y_a(k)-(N_a(k)-N_c(k))}$
$A_1 \rightarrow A_2 \wedge A_3 \wedge A_4$	$\frac{Y_d(k)}{N_d(k)}$	$\frac{Y_d(k)}{Y_a(k)-(N_a(k)-N_d(k))}$
$A_1 \wedge A_2 \rightarrow A_3$	$\frac{Y_c(k)}{N_c(k)}$	$\frac{Y_c(k)}{Y_b(k)-(N_b(k)-N_c(k))}$
$A_1 \wedge A_2 \rightarrow A_3 \wedge A_4$	$\frac{Y_d(k)}{N_d(k)}$	$\frac{Y_d(k)}{Y_b(k)-(N_b(k)-N_d(k))}$
$A_1 \wedge A_2 \wedge A_3 \rightarrow A_4$	$\frac{Y_d(k)}{N_d(k)}$	$\frac{Y_d(k)}{Y_c(k)-(N_c(k)-N_d(k))}$

$N_b(k) - N_d(k)$ is the number of tuples including missing data for $(A_3 = 1) \wedge (A_4 = 1)$ within $Y_b(k)$. Because the difference of the number of tuples including missing data between $(A_1 = 1) \wedge (A_2 = 1) \wedge (C = k)$ and $(A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1) \wedge (A_4 = 1) \wedge (C = k)$ equals to $N_b(k) - N_d(k)$.

The measurements of each rule for every class are calculated at the same time. Therefore, the rules showing the difference between classes in the database can be evaluated. Table 2 shows an example of measurements of rules in $C = k$. Using both measurements for $C = 1$ and $C = 0$, we can extract associative contrast rules.

In order to obtain the χ^2 value of the rules, we consider changes of the connection of Pointing nodes in each generation. For example, if the connection of P_1 is changed from '$A_1 = 1$' node to '$A_3 = 1$' node as shown in Fig. 3, we are able to calculate the support of $(A_3 = 1)$, $(A_3 = 1) \wedge (A_4 = 1)$ and $(A_3 = 1) \wedge (A_4 = 1) \wedge (A_5 = 1)$ in the next examination. In Fig. 3 (b), $Y_p(k)$, $Y_q(k)$ and $Y_r(k)$ are the numbers of tuples belonging to class k and moving to the Continue-side at each Judgment node satisfying that all the attribute values are equal to 1 from the Pointing node P_1. $M_p(k)$, $M_q(k)$ and $M_r(k)$ are also calculated at the same time. Then, the N values like $N_p(k)$, $N_q(k)$ and $N_r(k)$ are obtained using (8). When we calculate the χ^2 value of the rule $X \rightarrow Y$ in the incomplete database, we can use the N value of $X \cup Y$ instead of N in (1). α, β and γ in (1) are calculated by using Y values and N values. The operation changing the connections of the Pointing node can be repeated like a chain operation in each generation. A consequent of the rule can be the antecedent of another rule using this operation.

3.5 Extraction of Rules

In every generation, the examinations are done from $1 \in ID$ and P_1 node. Examinations of attribute values start from each Pointing node as described above. After all the tuples in the database are examined, measurements of candidate rules of every Pointing node are calculated and the interestingness of the rules are judged by given conditions. When an important rule is extracted, the overlap of the attributes is checked and it is also checked whether the important rule is new or not, i.e., whether it is in the pool or not. The extracted important rules are stored in a rule pool all together through the evolutionary process. Fig. 4 shows the flow of the rule extraction.

3.6 Genetic Operations and Fitness

Individuals are replaced with new ones by a selection rule in each generation [5]. The individuals are ranked by their fitnesses and upper 1/3 individuals are selected. The number 1/3 is determined experimentally, which is not so sensitive to the results. After that, they are reproduced three times for the next generation, then the following three kinds of genetic operators are executed to them; crossover with the probability of P_c, mutation-1 with the probability of P_{m1} (changes the connection of nodes) and mutation-2 with the probability of P_{m2} (changes the function of Judgment nodes). The operators are executed for the gene of Judgment nodes. All the connections of the Pointing nodes are changed randomly in order to extract new rules efficiently. $P_c = 1/5$, $P_{m1} = 1/3$ and $P_{m2} = 1/5$ is an effectual setting and was used in the experiments in Section 4. Information of the extracted rules like frequency of the appearances of attributes in the rules can be used for genetic operations. The more concrete explanation of the operations are described in [7].

Fitness of the individual can be defined depending on the problems. The capacity for extraction of new rules should be considered. Following functions were used in Section 4.

Fitness for associative contrast rule mining using χ^2 threshold is defined by

$$\begin{aligned} F_d^{\chi^2} &= \sum_{r \in R} \{ \chi^2_{(C=1)}(r) + 10(n_X(r) - 1) \\ &\quad + 10(n_Y(r) - 1) + \alpha_{new}(r) \} \end{aligned} \qquad (9)$$

where, R: set of suffixes of extracted rules satisfying (2), (3), (4) and (5) in the individual, $\chi^2_{(C=1)}(r)$: χ^2 value of rule r in $C = 1$. $n_X(r), n_Y(r)$: the number of attributes in the antecedent and in the consequent of rule r, respectively. $\alpha_{new}(r)$: additional constant defined by

$$\alpha_{new}(r) = \begin{cases} \alpha_{new} & \text{(rule } r \text{ is new)} \\ 0 & \text{(otherwise).} \end{cases} \qquad (10)$$

Constants are set up empirically. $\chi^2_{(C=1)}(r)$, $n_X(r)$, $n_Y(r)$ and $\alpha_{new}(r)$ are concerned with the importance, complexity and novelty of rule r, respectively.

Table 3: Averaged number of extracted rules (30 trials).

	missing rate (%)			
	0	2	5	10
GNP-based Method (Interesting rules) (Unexpected rules)	3450.5 (614.4) (0.0)	2356.4 (528.2) (400.6)	1269.8 (379.4) (397.6)	580.1 (147.8) (331.2)
Ring structure (Interesting rules) (Unexpected rules)	3333.2 (601.8) (0.0)	2297.1 (517.0) (382.2)	1250.2 (378.8) (384.7)	602.7 (156.6) (340.4)
Random network (Interesting rules) (Unexpected rules)	1416.9 (289.1) (0.0)	1117.2 (342.1) (208.1)	785.2 (260.8) (270.9)	491.0 (118.9) (296.1)

Fitness for using $confidence$ threshold is defined by

$$\begin{aligned} F_d^{conf} = & \textstyle\sum_{r \in R} \{ 10 \times |conf(r)_{(C=1)} - conf(r)_{(C=0)}| \\ & + (n_X(r) - 1) + (n_Y(r) - 1) + \alpha_{new}(r) \} \end{aligned} \quad (11)$$

where, R: set of suffixes of extracted rules satisfying (4), (5) and (6) or (7) in the individual, $conf(r)_{(C=k)}$: $confidence$ of rule r in $C = k$.

4. Experimental Results

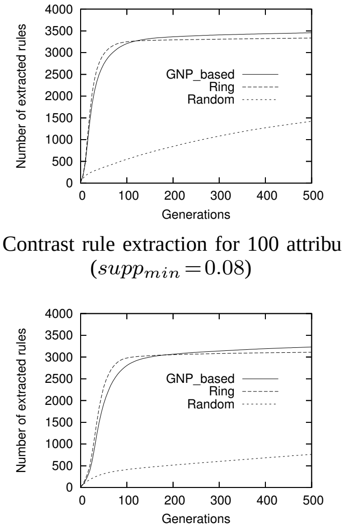
Experiments were executed using artificial incomplete data sets by the following viewpoints.

- Evaluation of the performance of the associative contrast rule extraction from the incomplete database.
- Evaluation of the mischief for the rule measurements by missing values.

We used the same dataset named SNP_{com} used in [5]. SNP_{com} has 100 attributes and 270 instances and has no missing data. The original data is The Mapping 500K HapMap Genotype Data Set (Affimetrix)[1]. This database contains Single Nucleotide Polymorphism (SNP) information of 270 people. 100 SNPs were picked up at random and constructed the dataset SNP_{com}. Support values of 100 SNPs are between 0.1 and 0.6. The original data has 4 class labels: YRI, JPT, CHB and CEU. Datasets including artificial missing values were generated randomly from SNP_{com} using given missing rates, i.e., 2%, 5% and 10%. For every missing rate, 30 incomplete data sets were generated and named $SNP_2(i)$, $SNP_5(i)$, and $SNP_{10}(i)$ ($i = 1, \ldots, 30$), respectively. In addition, we made a complete dataset having 200 attributes named as SNP_{com200} based on the above.

The population size for evolutionary rule accumulation mechanisms is 120. The number of Pointing nodes and Judgment nodes in each individual are 10 and 100, respectively. The number of changing the connections of the Pointing nodes in each generation is 5. The condition of termination is 500 generations for evolution. All algorithms were coded in C. Experiments were done on a 1.80GHz Intel(R) Core2 Duo CPU with 2GB RAM.

[1] http://www.affymetrix.com/support/technical/sample_data/500k_hapmap_genotype_data.affx



(a) Contrast rule extraction for 100 attributes. ($supp_{min} = 0.08$)

(b) Contrast rule extraction for 200 attributes. ($supp_{min} = 0.1$)

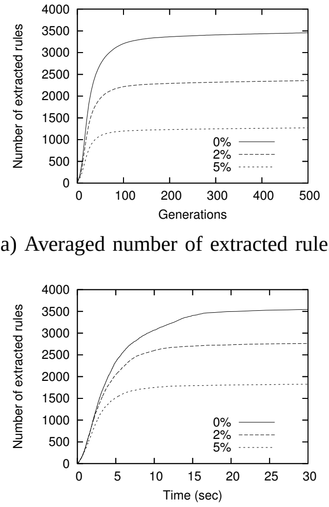
Fig. 5: Averaged number of extracted contrast rules.

First of all, the contrast rule mining in the SNP_{com} were evaluated. Instances were divided into 2 classes as follows; $C = 1$ in the case of YRI or JPT (135 instances), $C = 0$ in the case of CHB or CEU (135 instances). This class division has no scientific meaning, only intention was to make a dataset for the estimation use. The associative contrast rules defined by (2), (3), (4) and (5) were extracted. $supp_{min} = 0.08$, $\chi^2_{min} = 6.63$, $\chi^2_{max} = 1.0$, $1 \leq n_X(r) \leq 4$, $1 \leq n_Y(r) \leq 4$ and $\alpha_{new} = 150$ were used. In order to obtain the whole identified rules in the SNP_{com} satisfying the given conditions, 10000 independent rule extractions were done and obtained 4248 identified rules.

Fig. 5 (a) shows the averaged number of extracted rules over 30 data sets in the rule pool versus number of generations for the evolution. $GNP\text{-}based$, $Ring$ and $Random$ denote the methods described in Section 3. This demonstrates that the evolutionary rule accumulation based methods can extract most of the contrast rules within 100 generations. Fig. 5 (b) shows the same experiment in the case of using SNP_{com200}. In this experiment, $supp_{min} = 0.1$ was used. $Ring$ tends to converge in early generations.

Fig. 6 (a) shows the averaged number of extracted rules over 30 data sets. Associative contrast rules were extracted from SNP_{com} and $SNP_m(i)$ ($m = 2, 5$, $i = 1, \ldots, 30$), respectively. 0% denotes using SNP_{com}. 2% and 5% denote the missing rates. Fig. 6 (b) shows a sample of run-time in the same experiment as Fig. 6 (a). It shows that the most of the contrast rules were extracted within 10 seconds. In this experiment, 500 generations were set as the terminal condition, however, users can set the maximum calculation time instead and quit the rule extraction.

Table 3 shows the averaged number of total associative contrast rules obtained at the final generation. It is found that the method can extract rules based on χ^2 values from the dense incomplete database. The number of extracted rules tends to decrease by increasing the missing rate, this can be caused by the decrease of the N value in (1). In this experiment, *interesting rule* was defined as the rule extracted from SNP_{com} and satisfying additional conditions, that is, $\chi^2(X \rightarrow Y)_{(C=1)} \geq 10.0$, $support(X \rightarrow Y)_{(C=1)} \geq 0.1$ and $support(X \rightarrow Y)_{(C=0)} \geq 0.1$. The number of interesting rules in SNP_{com} is 642. It is found that 95% of the *interesting rules* are covered in each rule extraction using GNP-based method. In addition, *unexpected rule* was defined as the rule excluded from the rule extraction of SNP_{com}. A percentage of the number of *unexpected rules* tends to increase by the missing values.



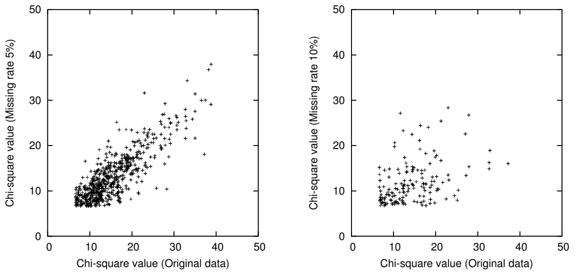
(a) Averaged number of extracted rules.

(b) Run-time versus number of extracted rules.

Fig. 6: Number of extracted contrast rules in the pool.

Fig. 7 (a) shows the scatter diagram of χ^2 values of extracted rules in $C = 1$ in the original data case and in the 5% missing rate case. Plots show the χ^2 values of all the rules obtained in the two rule extractions. 69% of the rules extracted in the 5% missing rate case are found in the rule pool of the original data case. It is found that most of the rules having high χ^2 value in the original data set are also extracted in the artificial incomplete data set using 5% missing rate. Fig. 7 (b) shows the scatter diagram for the 10% missing rate case. It shows the weak correlation of the χ^2 values of the rules compared with Fig. 7 (a). 43% of the rules extracted in the 10% missing rate case are found in the rule pool of the original data case. In this experiment, χ^2 values were used for the both classes as one of the conditions of interesting rules. This result suggests that 10% missing rate cause the different feature of rule extraction from the original data set in a detailed analysis.

Next, the associative contrast rule extraction between SNP_{com} and $SNP_m(i)$ ($m = 2, 5, 10$, $i = 1, \ldots, 30$) were examined based on (6) and (7) to evaluate the mischief for



(a) 5% missing rate case.

(b) 10% missing rate case.

Fig. 7: Scatter diagram of chi-square values.

the rule measurements by the missing rate. This experiment demonstrates the relationships between the missing rates and reliability of rule extraction. SNP_{com} is set at class $C=1$ and $SNP_m(i)$ is set at $C=0$. If many rules are extracted, then the missing values affects for the rule measurements, because $SNP_m(i)$ have different features from SNP_{com}. This experiment was executed using GNP-based method. $\delta = 0.03$, 0.05, 0.10 and 0.15 for (6) and (7) were used. The maximum number of extracted rules in the pool was set as 5000 and we quit the rule extraction by this condition. $1 \le n_X(r) \le 4$, $1 \le n_Y(r) \le 4$ and $\alpha_{new} = 30$ were used.

Table 4 shows the total number of extracted rules at 500 generation. ‘$-$’ describes that the number of extracted rule is more than 5000. In this experiment, a huge number of candidate rules are examined, however, only a small number of contrast rules are extracted in many cases. The associative contrast rule mining method can be used for the difference detection between two data sets.

Table 4: Averaged number of extracted contrast rules between original data and artificial incomplete data.

(a) $confidence(X \rightarrow Y)_{(original)}$
$-confidence(X \rightarrow Y)_{(artificial)} > \delta$

$supp_{min}$	δ	missing rate (%)		
		2	5	10
0.10	0.10	0.0	0.0	0.0
	0.05	0.2	0.5	0.0
	0.03	8.5	4.5	0.5
0.07	0.10	0.0	0.2	0.0
	0.05	4.4	3.0	0.4
	0.03	85.7	21.4	2.7
0.05	0.10	0.2	0.6	0.2
	0.05	23.7	13.9	3.1
	0.03	—	75.4	10.4
0.03	0.10	5.1	7.7	2.3
	0.05	—	104.7	22.6
	0.03	—	—	57.9
0.02	0.10	61.1	51.6	14.3
	0.05	—	—	76.4
	0.03	—	—	152.6

(b) $confidence(X \rightarrow Y)_{(artificial)}$
$-confidence(X \rightarrow Y)_{(original)} > \delta$

$supp_{min}$	δ	missing rate (%)		
		2	5	10
0.20	0.15	0.0	0.0	0.4
	0.10	0.0	0.2	25.3
	0.05	1.9	100.4	409.3
0.18	0.15	0.0	0.1	1.9
	0.10	0.0	2.2	63.1
	0.05	9.0	251.2	732.6
0.15	0.15	0.0	0.8	18.5
	0.10	0.0	17.1	374.3
	0.05	56.6	1242.6	2418.0
0.12	0.15	0.1	6.0	167.8
	0.10	1.2	134.7	2161.4
	0.05	419.6	—	—
0.10	0.15	0.2	28.0	855.0
	0.10	7.3	749.2	—
	0.05	1746.7	—	—

5. Conclusions

A method for associative contrast rule mining from incomplete databases has been demonstrated using a graph-based evolutionary method. An incomplete database includes missing data in some instances, however, the method can extract rules satisfying given conditions. The performances of the associative contrast rule extraction have been evaluated using artificial incomplete data sets in the medical field. The results show that the method has a potential to realize association analysis. In addition, the evaluation of the mischief for the rule measurements by missing values is demonstrated. We are studying applications of the method to information processing in the medical science field.

Acknowledgment.

This work was partly supported by JSPS KAKENHI Grant Number 24500191.

References

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, in *Proc. of the 20th VLDB Conf.*, pp. 487–499, 1994.

[2] J. Han, J. Pei, Y. Yin and R. Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach”, *Data Mining and Knowledge Discovery*, Vol. 8, pp. 53–87, 2004.

[3] J. W. Grzymala-Busse and W. J. Grzymala-Busse, Handling Missing Attribute Values Data Mining and Knowledge Discovery Handbook, 2nd ed., O. Maimon, L. Rockach (eds.), Springer, pp.33–51, 2010.

[4] M. Saar-Tsechansky and F. Provost, Handling Missing Values when Applying Classification Models, Journal of Machine Learning Research 8, pp.1625-1657, 2007.

[5] K. Shimada and K. Hirasawa, “A Method of Association Rule Analysis for Incomplete Database Using Genetic Network Programming”, in *Proc. of the Genetic and Evolutionary Computation Conference 2010 (GECCO 2010)* , pp. 1115–1122, 2010.

[6] K. Shimada, “An Evolving Associative Classifier for Incomplete Database”, Springer LNAI 7377: Advances in Data Mining, Perner P.(Ed.). pp.136–150, 2012.

[7] K. Shimada and K. Hirasawa, “Exceptional Association Rule Mining Using Genetic Network Programming”, in *Proc. of the 4th International Conference on Data Mining (DMIN 2008)*, pp. 277–283, 2008.

[8] S. Mabu, C. Chen, N. Lu, K. Shimada and K. Hirasawa, “An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming”, IEEE Trans. on Systems, Man, and Cybernetics - Part C-, Vol. 41, pp.130–139, 2011.

[9] A. A. Freitas, “Data Mining and knowledge Discovery with Evolutionary Algorithms”, Springer, 2002.

[10] A. Ghosh and L. C. Jain, “Evolutionary Computing in Data Mining”, Springer, 2005.