Families of the Granules for Association Rules and Their Properties

Hiroshi Sakai[1], Chenxi Liu[1], and Michinori Nakata[2]

[1] Graduate School of Engineering, Kyushu Institute of Technology,
Tobata, Kitakyushu 804-8550, Japan
sakai@mns.kyutech.ac.jp
[2] Faculty of Management and Information Science,
Josai International University
Gumyo, Togane, Chiba 283, Japan
nakatam@ieee.org

Abstract. We employed the granule (or the equivalence class) defined by a descriptor in tables, and investigated rough set-based rule generation. In this paper, we consider the new granules defined by an implication, and propose a *family of the granules defined by an implication* in a table with exact data. Each family consists of the four granules, and we show that three criterion values, *support*, *accuracy*, and *coverage*, can easily be obtained by using the four granules. Then, we extend this framework to tables with non-deterministic data. In this case, each family consists of the nine granules, and the minimum and the maximum values of three criteria are also obtained by using the nine granules. We prove that there is a table causing *support* and *accuracy* the minimum, and generally there is no table causing *support*, *accuracy*, and *coverage* the minimum. Finally, we consider the application of these properties to *Apriori*-based rule generation from uncertain data. These properties will make *Apriori*-based rule generation more effective.

Keywords: Association rules, Rule generation, Apriori algorithm, Granularity, Uncertainty.

1 Introduction

We coped with rule generation and data mining in *Non-deterministic Information Systems* (*NISs*) [8, 13, 15]. *NIS* was proposed by Pawlak [11], Orłowska [9], and Lipski [6, 7] in order to handle information incompleteness in the typical table defined as a *Deterministic Information System* (*DIS*) [10, 12, 17]. Pawlak's framework in *DIS* is called *Rough Set Theory*, and the equivalence classes take the important role. In [11], we see the definition of the *many valued system*, which is similar to *NIS*. Orłowska and Lipski considered question-answering methods in *NIS* independently.

We tried to extend rule generation in *DIS* to *NIS* by using the modal concepts [3], and proposed the framework *Rough Non-deterministic Information Analysis* (*RNIA*). In *RNIA*, we defined the *certain rules* and the *possible rules*, and proved

that the algorithm named *NIS-Apriori* is *sound* and *complete* for the defined rules [15, 16]. The *NIS-Apriori* algorithm is an adjusted *Apriori* algorithm [1, 2] to *NIS*. Even though *NIS-Apriori* handles the modal concepts in rules, the computational complexity is about twice the complexity of *Apriori*. Furthermore, we opened a web software tool *getRNIA* [18, 19].

In this paper, we propose the *families of the granules* in *NIS*, which are extended from the *division chart* [14]. In rough sets, we usually make use of the granules (or equivalence classes) defined by the descriptors. Some other types of granules are also proposed for handling missing values [5]. Here, we consider the granules defined by the implications. We have already coped with the six granules defined by the implications in [14], and we extend them to the nine granules. By this extension, we can consider the new criterion value *coverage* as well as *support* and *accuracy*.

This paper is organized as follows: Section 2 focuses on the case of the tables with exact data. We define the family of the four granules, and show the calculation of the criterion values. Section 3 considers the case of the tables with non-deterministic data. We similarly define the family of the nine granules, and show the extended results from Section 2. Section 4 describes the perspective of the *NIS-Apriori* algorithm based on the obtained results. Section 5 concludes this paper.

2 A Family of the Granules in DIS ψ

This section considers a family of the granules and its property in *DIS* ψ.

2.1 Preliminary

A *Deterministic Information System* (*DIS*) ψ is a quadruplet [10–12, 17]:

$$\psi = (OB, AT, \{VAL_A | \ A \in AT\}, f), \tag{1}$$

where OB is a finite set whose elements are called *objects*, AT is a finite set whose elements are called *attributes*, VAL_A is a finite set whose elements are called *attribute values* for an attribute $A \in AT$, and f is such a mapping:

$$f : OB \times AT \to \cup_{A \in AT} VAL_A. \tag{2}$$

We usually consider a table like Table 1 for ψ. A pair $[A, v]$ ($A \in AT$, $v \in VAL_A$) is called a *descriptor*, and we consider a set $CON \subseteq AT$ which we call (a set of) *condition attributes* and an attribute $Dec \in AT$ ($Dec \notin CON$) which we call a *decision attribute*. An *implication* τ for CON and Dec is a formula in the following:

$$\tau : \wedge_{A \in CON}[A, val_A] \Rightarrow [Dec, val] \ (val_A \in VAL_A, val \in VAL_{Dec}). \tag{3}$$

In most of work on rule generation, we try to obtain the appropriate implications, which we call rules. The most famous criterion for defining rules consists of three

Table 1. An exemplary deterministic information system ψ_1.

OB	$temperature$	$headache$	$nausea$	flu
1	$very\_high$	yes	yes	yes
2	$high$	yes	yes	yes
3	$normal$	yes	yes	no
4	$very\_high$	yes	no	yes
5	$very\_high$	yes	yes	yes
6	$high$	no	no	no
7	$normal$	no	yes	no
8	$high$	no	no	no

values, i.e., *support*, *accuracy* and *coverage* [10–12, 17]. We employ these values, and we define that a *rule* is an implication τ satisfying the constraint

$$\begin{array}{l} support(\tau) \geq \alpha, accuracy(\tau) \geq \beta \text{ and } coverage(\tau) \geq \gamma \\ \text{for given } 0 < \alpha, \beta, \gamma \leq 1. \end{array} \tag{4}$$

The constraint *coverage* may not be employed in some frameworks, for example the *Apriori* algorithm [1, 2] does not employ it, and we see γ=0 in such case. In ψ_1, we have $support(\tau)$=3/8, $accuracy(\tau)$=3/4, and $coverage(\tau)$=3/4 for $\tau : [headache, yes] \wedge [nausea, yes] \Rightarrow [flu, yes]$.

2.2 A Family of the Granules Defined by an Implication

We propose the granules defined by an implication $\tau : \wedge_{A \in CON}[A, val_A] \Rightarrow [Dec, val]$.

Definition 1. *We say an object x supports τ, if $f(x, A)$=val_A for every $A \in CON$ and $f(x, Dec)$=val. In order to clarify this object x, we may employ the notation τ^x instead of τ.*

In ψ_1, the objects 1, 2 and 5 support $\tau : [headache, yes] \wedge [nausea, yes] \Rightarrow [flu, yes]$, and we have τ^1, τ^2, and τ^5.

For more simplicity of $\tau : \wedge_{A \in CON}[A, val_A] \Rightarrow [Dec, val]$, we employ the following notation:
(1) p denotes the conjunction $\wedge_{A \in CON}[A, val_A]$,
(2) p' denotes any conjunction $\wedge_{A \in CON}[A, val'_A]$ ($val'_A \neq val_A$ for at least one $A \in CON$),
(3) r denotes $[Dec, val]$,
(4) r' denotes any descriptor $[Dec, val']$ ($val' \neq val$).
If we fix an implication $\tau : p \Rightarrow r$, each object defines either $p \Rightarrow r$, $p \Rightarrow r'$, $p' \Rightarrow r$, or $p' \Rightarrow r'$. In ψ_1, let us consider $\tau : [headache, yes] \wedge [nausea, yes] \Rightarrow [flu, yes]$. Then, the objects 1, 2, and 5 define $p \Rightarrow r$, the object 3 defines $p \Rightarrow r'$, the object 4 does $p' \Rightarrow r$, and the objects 6, 7, and 8 do $p' \Rightarrow r'$.

Table 2. Four granules defined by $\tau : p \Rightarrow r$ in ψ.

	r	r'
p	①=$\{x \in OB \mid x \text{ supports } p \Rightarrow r\}$	②=$\{x \in OB \mid x \text{ supports } p \Rightarrow r'\}$
p'	③=$\{x \in OB \mid x \text{ supports } p' \Rightarrow r\}$	④=$\{x \in OB \mid x \text{ supports } p' \Rightarrow r'\}$

Based on these four types of implications, we define four sets ①, ②, ③, and ④ in Table 2. Since we can show the following,

$$① \cup ② \cup ③ \cup ④ = OB, \ ① \cap ⓙ = \emptyset \ (i \neq j), \tag{5}$$

the four sets are equivalence classes over OB.

Definition 2. *For DIS ψ, an implication τ and the four equivalence classes in Table 2, we define $FGr(\tau, \psi)$=(①, ②, ③, ④), and we say $FGr(\tau, \psi)$ is a family of the granules defined by τ in ψ.*

In our previous work, we proposed a set of equivalence classes and named it a *division chart* [14]. In a division chart, we handled two types of implications, namely $p \Rightarrow r$ and $p \Rightarrow r'$, and considered only two granules ① and ②. We calculated *support* and *accuracy* by using ① and ②, however we can also calculate *coverage* by using $FGr(\tau, \psi)$. Thus, we are extending the previous work [14] to the more powerful one. Since $FGr(\tau, \psi)$ takes the role of the contingency table, it is easy to obtain the following proposition.

Proposition 1. *For a family $FGr(\tau, \psi)$ of the granules, the following holds.*

$$\begin{aligned} &(1)\ support(\tau) = \frac{|①|}{|①|+|②|+|③|+|④|} = \frac{|①|}{|OB|}, \\ &(2)\ accuracy(\tau) = \frac{|①|}{|①|+|②|}, \quad (3)\ coverage(\tau) = \frac{|①|}{|①|+|③|}. \end{aligned} \tag{6}$$

Remark 1. For $\tau : [temperature, normal] \Rightarrow [flu, no]$ $(=p \Rightarrow r)$ in ψ_1, we have $FGr(\tau, \psi_1)$=$(\{3,7\}, \emptyset, \{6,8\}, \{1,2,4,5\})$. Based on Proposition 1,

$$\begin{aligned} &(1)\ support(\tau) = \frac{|\{3,7\}|}{8} = 1/4, \ (2)\ accuracy(\tau) = \frac{|\{3,7\}|}{|\{3,7\}|+|\emptyset|} = 1.0, \\ &(3)\ coverage(\tau) = \frac{|\{3,7\}|}{|\{3,7\}|+|\{6,8\}|} = 1/2. \end{aligned} \tag{7}$$

Especially, $1, 2 \in ④$ seems quite new, as far as authors know. In the typical equivalence relation, 1 and 2 are not in the same class, however they are equivalent in the aspect that neither object 1 nor 2 is related to the implication τ at all.

3 A Family of the Granules in NIS Φ

We have dealt with *NIS* as the case of the tables with uncertainty. In this section, we consider a family of the granules in *NIS* and their property.

Table 3. An exemplary non-deterministic information system Φ_1.

OB	$temperature$	$headache$	$nausea$	flu
1	$\{very\_high\}$	$\{yes, no\}$	$\{yes\}$	$\{yes\}$
2	$\{high, very\_high\}$	$\{yes\}$	$\{yes\}$	$\{yes\}$
3	$\{normal, high\}$	$\{yes\}$	$\{yes\}$	$\{yes, no\}$
4	$\{very\_high\}$	$\{yes\}$	$\{yes, no\}$	$\{yes\}$
5	$\{very\_high\}$	$\{yes, no\}$	$\{yes\}$	$\{yes\}$
6	$\{high\}$	$\{no\}$	$\{yes, no\}$	$\{yes, no\}$
7	$\{normal\}$	$\{no\}$	$\{yes\}$	$\{no\}$
8	$\{normal, high\}$	$\{no\}$	$\{yes, no\}$	$\{no\}$

3.1 Preliminary

NIS Φ is also a quadruplet $\Phi=(OB, AT, \{VAL_A | A \in AT\}, g)$ [9, 11], where g is such a mapping:

$$g : OB \times AT \to P(\cup_{A \in AT} VAL_A) \text{ (a power set of } \cup_{A \in AT} VAL_A). \quad (8)$$

Every set $g(x, A)$ is interpreted as that there is an actual value in this set, but the value is not known. We usually consider tabular representation of Φ like Table 3. Since each VAL_A is a finite set, we can easily define all possible cases from *NIS*. For $\Phi=(OB, AT, \{VAL_A | A \in AT\}, g)$, we name the following *DIS* a *derived DIS* from *NIS* Φ.

$$\psi = (OB, AT, \{VAL_A | A \in AT\}, h) \ (h(x, A) \in g(x, A) \text{ for each } x \text{ and } A). \quad (9)$$

For *NIS* Φ, let $DD(\Phi)$ denote a set below:

$$DD(\Phi) = \{\psi \mid \psi \text{ is a derived } DIS \text{ from } \Phi\}. \quad (10)$$

Actually, ψ_1 in Table 1 is a derived *DIS* from Φ_1. We transfer the problem on information incompleteness to the case analytic problem based on $DD(\Phi)$. We define that τ is a *certain rule*, if τ is a rule in every $\psi \in DD(\Phi)$, and τ is a *possible rule*, if τ is a rule in at least one $\psi \in DD(\Phi)$. However, there are 1024 $(=2^{10})$ derived *DISs* in Φ_1. In Mammographic data set [4], there are more than 10^{100} derived *DISs*. For solving this problem, the family of the granules takes the important role.

Definition 3. *We say an object x supports τ in NIS Φ, if x supports τ in at least one $\psi \in \Phi$.*

Remark 2. For $\tau : [temperature, normal] \Rightarrow [flu, no]$ in Φ_1, we have τ^3, τ^7, and τ^8, and we say each of them is an *instance* of τ. In order to evaluate τ in *NIS* Φ, we consider the instance of τ in Φ. If there is an instance τ^x satisfying the constraint, we say τ is a rule. In *DIS*, we need not to consider such instances, because three criterion values are the same for every τ^x and τ^y $(x \neq y)$. However in *NIS*, they may be different. For example, the instance τ^7 occurs in each of the derived *DIS*, but there is a derived *DIS* where τ^3 nor τ^8 do not occur. They have the different property from τ^7.

3.2 A Family of the Granules in NIS Φ

For Φ, an implication τ, and an object $x \in OB$, we define the following:

$$\begin{array}{l} IMP(x,\Phi,\tau) = \cup_{\psi \in DD(\Phi)}\{\tau_\psi \mid x \text{ supports } \tau_\psi \text{ in } DIS\ \psi\}, \\ \text{Here, } \tau_\psi \text{ is either } \tau : p \Rightarrow r,\ p \Rightarrow r',\ p' \Rightarrow r \text{ or } p' \Rightarrow r'. \end{array} \tag{11}$$

This $IMP(x,\Phi,\tau)$ means a set of the obtainable implications, which are classified by τ, from x. For example in Φ_1, we consider τ : $[temperature, normal] \Rightarrow [flu, no]$ $(= p \Rightarrow q)$. Then, we have the following:

$$\begin{array}{l} IMP(3,\Phi_1,\tau) = \{p \Rightarrow r, p \Rightarrow r', p' \Rightarrow r, p' \Rightarrow r'\}, \\ IMP(7,\Phi_1,\tau) = \{p \Rightarrow r\}, \\ IMP(8,\Phi_1,\tau) = \{p \Rightarrow r, p' \Rightarrow r\}. \end{array} \tag{12}$$

Proposition 2. *The relation $\sim_{\tau,\Phi}$ below defines an equivalence relation over OB.*

$$x \sim_{\tau,\Phi} y \Leftrightarrow IMP(x,\Phi,\tau) = IMP(y,\Phi,\tau). \tag{13}$$

Proof. *We can easily show the reflexivity, the symmetry and the transitivity.*

For p $(=\wedge_{A \in CON}[A, val_A])$ in τ, it is necessary to consider three cases, i.e., $\{p\}$, $\{p, p'\}$ and $\{p'\}$. Similarly, we think three cases, $\{r\}$, $\{r, r'\}$ and $\{r'\}$, and we have the nine sets based on the obtainable implications in Table 4.

Table 4. Nine sets based on the obtainable implications for $\tau : p \Rightarrow r$.

	$\{r\}$	$\{r, r'\}$	$\{r'\}$
$\{p\}$	$S_1 : \{p \Rightarrow r\}$	$S_2 : \{p \Rightarrow r, p \Rightarrow r'\}$	$S_3 : \{p \Rightarrow r'\}$
$\{p, p'\}$	$S_4 : \{p \Rightarrow r,$ $p' \Rightarrow r\}$	$S_5 : \{p \Rightarrow r, p \Rightarrow r',$ $p' \Rightarrow r, p' \Rightarrow r'\}$	$S_6 : \{p \Rightarrow r',$ $p' \Rightarrow r'\}$
$\{p'\}$	$S_7 : \{p' \Rightarrow r\}$	$S_8 : \{p' \Rightarrow r, p' \Rightarrow r'\}$	$S_9 : \{p' \Rightarrow r'\}$

Definition 4. *We define the following based on Table4.*

$$\begin{array}{l} ⓘ = \{x \in OB \mid IMP(x,\Phi,\tau) = S_i\}\ (1 \leq i \leq 9), \\ FGr(\tau,\Phi) = (①, ②, ③, ④, ⑤, ⑥, ⑦, ⑧, ⑨). \end{array} \tag{14}$$

We say $FGr(\tau,\Phi)$ is a family of the granules defined by τ in Φ.

For example, we have the following for τ : $[temperature, normal] \Rightarrow [flu, no]$ in Φ_1 and the formulas (12).

$$\begin{array}{l} IMP(3,\Phi_1,\tau) = S_5 \text{ and } 3 \in ⑤, \\ IMP(7,\Phi_1,\tau) = S_1 \text{ and } 7 \in ①, \\ IMP(8,\Phi_1,\tau) = S_4 \text{ and } 8 \in ④, \\ FGr(\tau,\Phi_1) = (\{7\}, \emptyset, \emptyset, \{8\}, \{3\}, \emptyset, \emptyset, \{6\}, \{1,2,4,5\}). \end{array} \tag{15}$$

3.3 Criterion Values of an Implication in NIS Φ

In *NIS* Φ, the criterion values depend upon $\psi \in DD(\Phi)$, so we consider the minimum and the maximum values of $support(\tau^x)$, $accuracy(\tau^x)$, and $coverage(\tau^x)$. We employ the notations $minsupp(\tau^x)$, $maxsupp(\tau^x)$, $minacc(\tau^x)$, $maxacc(\tau^x)$, $mincov(\tau^x)$, $maxcov(\tau^x)$ for them. We say τ^x is *definite* in Φ, if $IMP(x, \Phi, \tau)$ is a singleton set. If τ^x is not definite, there is at least on $\psi \in DD(\Phi)$ where τ^x does not occur. In this ψ, we define $minsupp(\tau^x)$=$minacc(\tau^x)$= $mincov(\tau^x)$=0. Now, we sequentially consider three criterion values by generating an actual $\psi \in DD(\Phi)$.

Proposition 3. *For NIS* Φ *and an implication* τ*, let us consider* $FGr(\tau, \Phi)$ $=(①, ②, ③, ④, ⑤, ⑥, ⑦, ⑧, ⑨)$*. If* $① \neq \emptyset$*, there is an object* $x \in ①$ *and* $\psi_{minSA} \in DD(\Phi)$ *satisfying the following:*

$$\begin{aligned}
&(1)\ support(\tau^x)\ in\ \psi_{minSA} = minsupp(\tau^x) = |①|/|OB|,\\
&(2)\ accuracy(\tau^x)\ in\ \psi_{minSA} = minacc(\tau^x) = \frac{|①|}{|①|+|②|+|③|+|⑤|+|⑥|},\\
&(3)\ coverage(\tau^x)\ in\ \psi_{minSA} = \frac{|①|}{|①|+|④|+|⑦|+|⑧|}.
\end{aligned} \tag{16}$$

(*Sketch of the proof*) *By the selection of an implication from the sets* S_2*,* S_4*,* S_5*,* S_6*,* S_8*, the criterion values change. For two natural number* N *and* M $(N \leq M)$*, we can easily show the inequality* $\frac{N}{M} \leq \frac{N+1}{M+1}$*. If we select* $\tau^x : p \Rightarrow r$ *in* S_2*, this* x *satisfies the condition on the denominator and the numerator of accuracy. Based on the inequality, this selection causes to increase accuracy. Thus, we select* $p \Rightarrow r'$*. Namely, we employ the strategy to select the same condition and the different decision. At the same time, we implicitly specify a table by this selection. If we select the underlined part in Table 5, we have* ψ_{minSA}*, and both support and accuracy are the minimum. If* $①=\emptyset$*, we handle* τ^x $(x \in ②\cup④\cup⑤)$*. In this case, we also have the formulas for the calculation, however the formulas are slightly different. We omit this case.*

Table 5. The selection (underlined part) of the implications for $\psi_{minSA} \in DD(\Phi)$.

	$\{r\}$	$\{r, r'\}$	$\{r'\}$
$\{p\}$	$S_1 : \{p \Rightarrow r\}$	$S_2 : \{p \Rightarrow r, \underline{p \Rightarrow r'}\}$	$S_3 : \{p \Rightarrow r'\}$
$\{p, p'\}$	$S_4 : \{p \Rightarrow r,$ $\underline{p' \Rightarrow r}\}$	$S_5 : \{p \Rightarrow r, \underline{p \Rightarrow r'},$ $p' \Rightarrow r, p' \Rightarrow r'\}$	$S_6 : \{\underline{p \Rightarrow r'},$ $\underline{p' \Rightarrow r'}\}$
$\{p'\}$	$S_7 : \{\underline{p' \Rightarrow r}\}$	$S_8 : \{\underline{p' \Rightarrow r}, p' \Rightarrow r'\}$	$S_9 : \{p' \Rightarrow r'\}$

Example 1. For $\tau : [temperature, normal] \Rightarrow [flu, no]$, we obtained $FGr(\tau, \Phi_1)$= $(\{7\}, \emptyset, \emptyset, \{8\}, \{3\}, \emptyset, \emptyset, \{6\}, \{1, 2, 4, 5\})$ in the formulas (15). Since $7 \in ①$, we can apply Proposition 3 to τ^7. Based on Table 5, we select $[temperature, high] \Rightarrow [flu, no]$ from object 8 and $[temperature, normal] \Rightarrow [flu, yes]$ from object 3. Then, we have ψ_{minSA} in Table 6. In ψ_{minSA}, $support(\tau^7)$=1/8=$minsupp(\tau^7)$, $accuracy(\tau^7)$=1/2=$minacc(\tau^7)$, $coverage(\tau^7)$=1/3>$mincov(\tau^7)$.

Table 6. ψ_{minSA} from $DD(\Phi_1)$.

OB	$temperature$	flu
1	$very\_high$	yes
2	$high$	yes
3	$normal$	yes
4	$very\_high$	yes
5	$very\_high$	yes
6	$high$	no
7	$normal$	no
8	$high$	no

Table 7. ψ_{minSC} from $DD(\Phi_1)$.

OB	$temperature$	flu
1	$very\_high$	yes
2	$high$	yes
3	$high$	no
4	$very\_high$	yes
5	$very\_high$	yes
6	$high$	no
7	$normal$	no
8	$high$	no

Proposition 4. *For NIS Φ and an implication τ, let us consider $FGr(\tau,\Phi)$ $=(①,②,③,④,⑤,⑥,⑦,⑧,⑨)$. If $① \neq \emptyset$, there is an object $x \in ①$ and $\psi_{minSC} \in DD(\Phi)$ satisfying the following:*

$$\begin{array}{ll} (1)\ support(\tau^x)\ in\ \psi_{minSC} = minsupp(\tau^x) & = |①|/|OB|, \\ (2)\ accuracy(\tau^x)\ in\ \psi_{minSC} = \dfrac{|①|}{|①|+|②|+|③|+|⑥|}, & \\ (3)\ coverage(\tau^x)\ in\ \psi_{minSC} = mincov(\tau^x) & = \dfrac{|①|}{|①|+|④|+|⑤|+|⑦|+|⑧|}. \end{array} \tag{17}$$

(*Sketch of the proof*) *We similarly have the selection in Table 8, which defines the minimum value of coverage. At the same time, ψ_{minSC} defines $minsupp(\tau^x)$, because $p \Rightarrow r$ occurs only in S_1.*

Table 8. The selection (underlined part) of the implications for $\psi_{minSC} \in DD(\Phi)$.

	$\{r\}$	$\{r,r'\}$	$\{r'\}$
$\{p\}$	$S_1:\{p \Rightarrow r\}$	$S_2:\{p \Rightarrow r, \underline{p \Rightarrow r'}\}$	$S_3:\{p \Rightarrow r'\}$
$\{p,p'\}$	$S_4:\{p \Rightarrow r,$ $\underline{p' \Rightarrow r}\}$	$S_5:\{p \Rightarrow r, p \Rightarrow r',$ $\underline{p' \Rightarrow r}, p' \Rightarrow r'\}$	$S_6:\{\underline{p \Rightarrow r'},$ $\underline{p' \Rightarrow r'}\}$
$\{p'\}$	$S_7:\{p' \Rightarrow r\}$	$S_8:\{\underline{p' \Rightarrow r}, p' \Rightarrow r'\}$	$S_9:\{p' \Rightarrow r'\}$

Example 2. In the same condition in Example 1, we apply Proposition 4 to the instance τ^7. Then, we have $support(\tau^7)$=1/8=$minsupp(\tau^7)$, $accuracy(\tau^7)$=1.0> $minacc(\tau^7)$, $coverage(\tau^7)$=1/4=$mincov(\tau^7)$. As the side effect, we obtain a derived ψ_{minSC} in Table 7.

Proposition 5. *For NIS Φ and an implication τ, let us consider $FGr(\tau,\Phi)$ $=(①,②,③,④,⑤,⑥,⑦,⑧,⑨)$, and let us suppose $① \neq \emptyset$. There is an object $x \in ①$ and $\psi_{minSAC} \in DD(\Phi)$ satisfying the following, if and only if $⑤=\emptyset$.*

$$\begin{array}{ll} (1)\ support(\tau^x)\ in\ \psi_{minSAC} = minsupp(\tau^x) = |①|/|OB|, & \\ (2)\ accuracy(\tau^x)\ in\ \psi_{minSAC} = minacc(\tau^x) = \dfrac{|①|}{|①|+|②|+|③|+|⑥|}, & \\ (3)\ coverage(\tau^x)\ in\ \psi_{minSAC} = mincov(\tau^x) & = \dfrac{|①|}{|①|+|④|+|⑦|+|⑧|}. \end{array} \tag{18}$$

(*Sketch of the proof*) *Based on Table 5 and 8, only the selection in* ⑤ *is different. If* ⑤=$\emptyset$, *the selections in Table 5 and 8 are the same.*

We generally know there is no $\psi_{minSAC} \in DD(\Phi)$ defining $minsupp(\tau^x)$, $minacc(\tau^x)$, and $mincov(\tau^x)$ based on Proposition 5. However, there are ψ_{minSA} defining $minsupp(\tau^x)$ and $minacc(\tau^x)$, and ψ_{minSC} defining $minsupp(\tau^x)$ and $mincov(\tau^x)$. Namely, we recognize $minsupp(\tau^x)$, $minacc(\tau^x)$, and $mincov(\tau^x)$ by examining at most two derived *DISs* ψ_{minSA} and ψ_{minSC}. Now, we consider the maximum case. We have the following.

Proposition 6. *For NIS* Φ *and an implication* τ, *let us consider* $FGr(\tau, \Phi)$ $=(①, ②, ③, ④, ⑤, ⑥, ⑦, ⑧, ⑨)$. *For any object* $x \in ① \cup ② \cup ④ \cup ⑤$, *there is* $\psi_{maxSAC} \in DD(\Phi)$ *satisfying the following:*

$$\begin{aligned}
&(1)\ support(\tau^x)\ in\ \psi_{maxSAC} = maxsupp(\tau^x) = \frac{(|①|+|②|+|④|+|⑤|)}{|OB|},\\
&(2)\ accuracy(\tau^x)\ in\ \psi_{maxSAC} = maxacc(\tau^x) = \frac{(|①|+|②|+|④|+|⑤|)}{(|①|+|②|+|③|+|④|+|⑤|)},\\
&(3)\ coverage(\tau^x)\ in\ \psi_{maxSAC} = maxcov(\tau^x) = \frac{(|①|+|②|+|④|+|⑤|)}{(|①|+|②|+|④|+|⑤|+|⑦|)}.
\end{aligned} \tag{19}$$

(*Sketch of the proof*) *Based on Table 9, we can similarly show the equations.*

Table 9. The selection (underlined part) of the implications for $\psi_{maxSAC} \in DD(\Phi)$.

	$\{r\}$	$\{r, r'\}$	$\{r'\}$
$\{p\}$	$S_1 : \{p \Rightarrow r\}$	$S_2 : \{\underline{p \Rightarrow r}, p \Rightarrow r'\}$	$S_3 : \{p \Rightarrow r'\}$
$\{p, p'\}$	$S_4 : \{\underline{p \Rightarrow r}, p' \Rightarrow r\}$	$S_5 : \{\underline{p \Rightarrow r}, p \Rightarrow r', p' \Rightarrow r, p' \Rightarrow r'\}$	$S_6 : \{p \Rightarrow r', \underline{p' \Rightarrow r'}\}$
$\{p'\}$	$S_7 : \{p' \Rightarrow r\}$	$S_8 : \{p' \Rightarrow r, \underline{p' \Rightarrow r'}\}$	$S_9 : \{p' \Rightarrow r'\}$

4 Criterion Values and Apriori-based Rule Generation

This section applies the obtained results to *Apriori*-based rule generation.

4.1 Current Rule Generation by Criteria *support* and *accuracy*

By Proposition 3, there is a derived $\psi_{minSA} \in DD(\Phi)$ for τ^x ($x \in ①$), and we can prove (C1) (the definition of a certain rule by *support* and *accuracy*) and (C2) are equivalent [15].

(C1) $support(\tau^x) \geq \alpha$ and $accuracy(\tau^x) \geq \beta$ in each $\psi \in DD(\Phi)$,

(C2) $support(\tau^x) \geq \alpha$ and $accuracy(\tau^x) \geq \beta$ in ψ_{minSA}, namely $minsupp(\tau^x) \geq \alpha$ and $minacc(\tau^x) \geq \beta$.

Similarly, we can prove (P1) (the definition of a possible rule by *support* and *accuracy*) and (P2) are equivalent by Proposition 6.
(P1) $support(\tau^x) \geq \alpha$ and $accuracy(\tau^x) \geq \beta$ in at least one $\psi \in DD(\Phi)$,
(P2) $support(\tau^x) \geq \alpha$ and $accuracy(\tau^x) \geq \beta$ in ψ_{maxSAC}, namely
$maxsupp(\tau^x) \geq \alpha$ and $maxacc(\tau^x) \geq \beta$.

Even though the conditions (C1) and (P1) depend upon $|DD(\Phi)|$, the conditions (C2) and (P2) do not depend upon $|DD(\Phi)|$. We can calculate the conditions (C2) and (P2) in the polynomial time order, and we escaped from the computational complexity problem on $|DD(\Phi)|$.

We have opened a software *getRNIA* powered by *NIS-Apriori* algorithm [18, 19]. In this implementation, we handled *support* and *accuracy*, and did not handle *coverage*. The *getRNIA* actually calculates the conditions (C2) and (P2) instead of the definitions (C1) and (P1), respectively. Moreover, *getRNIA* employs the merging procedure internally. For two families $FGr((p_1 \Rightarrow r), \Phi)$ and $FGr((p_2 \Rightarrow r), \Phi)$, we can obtain $FGr((p_1 \wedge p_2 \Rightarrow r), \Phi)$ [14]. After merging them, we can similarly apply Proposition 3 to 6.

4.2 Rule Generation by Criteria *support*, *accuracy*, and *coverage*

At first, we consider *Apriori*-based possible rule generation. By Proposition 6, there is a derived $\psi_{maxSAC} \in DD(\Phi)$ for any τ^x, so we can easily prove the conditions (P'1) (the definition of a certain rule by *support*, *accuracy*, and *coverage*) and (P'2) are equivalent.
(P'1) $support(\tau^x) \geq \alpha$, $accuracy(\tau^x) \geq \beta$, and $coverage(\tau^x) \geq \gamma$
in at least one $\psi \in DD(\Phi)$,
(P'2) $support(\tau^x) \geq \alpha$, $accuracy(\tau^x) \geq \beta$, and $coverage(\tau^x) \geq \gamma$ in ψ_{maxSAC},
namely $maxsupp(\tau^x) \geq \alpha$, $maxacc(\tau^x) \geq \beta$, and $maxcov(\tau^x) \geq \gamma$.
Therefore, we employ the condition (P'2) for possible rule generation. The following is an overview of *Apriori*-based possible rule generation.

An overview of Apriori-based possible rule generation in *NIS*
(Base step)
Set i=1. Prepare $LIST_i$:=$\{\tau : [A, val_A] \Rightarrow [Dec, val]\}$, and ANS:=$\{\}$.
(Inductive step)
Generate $FGr(\tau, \Phi)$ for each $\tau \in LIST_i$ by searching the total data set, and examine the following:
(1) $REST$:=$\{\}$. Set i:=i+1.
(2) Apply Proposition 6 for every $FGr(\tau, \Phi)$.
If τ^x satisfies the constraint, $ANS := ANS \cup \{\tau\}$. If $maxsupp(\tau^x) \geq \alpha$, $maxcov(\tau^x) \geq \gamma$, and $maxacc(\tau^x) < \beta$, $REST := REST \cup \{\tau\}$.
(3) $LIST_i$:=$\{\}$. For $\tau_j : con_j \Rightarrow r, \tau_k : con_k \Rightarrow r \in REST$,
generate $\tau : condition \Rightarrow r$ (the condition is a conjunction consisting of i number of descriptors), and $LIST_i$:=$LIST_i \cup \{\tau\}$.
(4) If $LIST_i$ is an empty set, this program terminates. All certain rules are stored in ANS. Otherwise, repeat the inductive step.

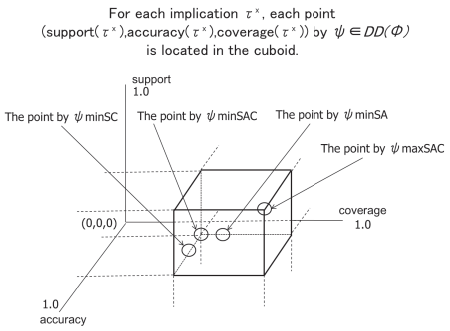

Fig. 1. The relation between ψ_{minSA}, ψ_{minSC}, ψ_{minSAC}, and ψ_{maxSAC}.

Now, let us consider certain rule generation based on Proposition 3, 4, and 5. For τ^x ($x \in$ ①), the conditions (C'1) and (C'2) are equivalent, if the granule ⑤=$\emptyset$ in $FGr(\tau, \Phi)$.

(C'1) $support(\tau^x) \geq \alpha$, $accuracy(\tau^x) \geq \beta$, and $coverage(\tau^x) \geq \gamma$ in each $\psi \in DD(\Phi)$,

(C'2) $support(\tau^x) \geq \alpha$, $accuracy(\tau^x) \geq \beta$, and $coverage(\tau^x) \geq \gamma$ in ψ_{minSAC}, namely $minsupp(\tau^x) \geq \alpha$, $minacc(\tau^x) \geq \beta$, and $mincov(\tau^x) \geq \gamma$.

If ⑤ $\neq \emptyset$, we need to consider the condition (C'3).

(C'3) $support(\tau^x) \geq \alpha$, $accuracy(\tau^x) \geq \beta$ in ψ_{minSA}, and $coverage(\tau^x) \geq \gamma$ in ψ_{minSC}.

Figure 1 shows the survey. We will employ the condition (C'3) for adding the criterion *coverage* to the *NIS-Apriori* algorithm as well as *support* and *accuracy*.

5 Concluding Remarks

This paper proposed a family of the granules $FGr(\tau, \Phi)$, and examined its property related to rule generation and data mining. We showed the calculation on $mincov(\tau^x)$ and $maxcov(\tau^x)$, and proved that we always have $\psi_{minSA} \in DD(\Phi)$, $\psi_{minSC} \in DD(\Phi)$, and $\psi_{maxSAC} \in DD(\Phi)$. As for $\psi_{minSAC} \in DD(\Phi)$, generally we may not have it. We proved the necessary and sufficient condition for existing $\psi_{minSAC} \in DD(\Phi)$. The computational complexity for calculating criterion values depends upon $DD(\Phi)$ in the definition, however we can calculate them in the polynomial time based on the properties of $FGr(\tau, \Phi)$. The content in this paper will be the mathematical foundation on $FGr(\tau, \Phi)$, and such background will enhance *Apriori*-based rule generation.

Acknowledgment: The authors would be grateful for reviewers' useful comments. This work is supported by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant Number 26330277.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. *Proc. VLDB'94*, Morgan Kaufmann, 487–499 (1994)
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. in: *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 307–328 (1996)
3. Blackburn, P. et al.: Modal Logic. Cambridge University Press (2001)
4. Frank, A., Asuncion, A.: UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science (2010) `http://mlearn.ics.uci.edu/MLRepository.html`
5. J.W. Grzymała-Busse, Data with missing attribute values: Generalization of indiscernibility relation and rule induction, *Transactions on Rough Sets* 1 (2004) 78–95.
6. Lipski, W.: On semantic issues connected with incomplete information databases. *ACM Transactions on Database Systems* 4(3), 262–296 (1979)
7. Lipski, W.: On databases with incomplete information. *Journal of the ACM* 28(1), 41–70 (1981)
8. Nakata, M., Sakai, H.: Twofold rough approximations under incomplete information. *Int'l. J. General Systems* 42(6), 546–571 (2013)
9. Orłowska, E., Pawlak, Z.: Representation of nondeterministic information. *Theoretical Computer Science* 29(1-2), 27–39 (1984)
10. Pawlak, Z.: Information systems theoretical foundations. *Information Systems* 6(3), 205–218 (1981)
11. Pawlak, Z.: Systemy Informacyjne: Podstawy Teoretyczne (in Polish) WNT (1983)
12. Pawlak, Z.: Rough Sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers (1991)
13. Sakai, H., Ishibashi, R., Koba, K., Nakata, M.: Rules and apriori algorithm in non-deterministic information systems. *Transactions on Rough Sets* 9, 328–350 (2008)
14. Sakai, H., Wu, M., Nakata, M.: Division charts as granules and their merging algorithm for rule generation in nondeterministic data. *Int'l. J. Intelligent Systems* 28(9), 865–882 (2013)
15. Sakai, H., Wu, M., Nakata, M.: Apriori-based rule generation in incomplete information databases and non-deterministic information systems. *Fundamenta Informaticae* 130(3), 343–376 (2014)
16. Sakai, H., Wu, M., Nakata, M.: The completeness of NIS-Apriori algorithm and a software tool getRNIA. *Proc. Int'l. Conf on AAI2014*, 115-121 (2014)
17. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. in: R. Słowiński (ed.), Intelligent Decision Support - Handbook of Advances and Applications of the Rough Set Theory, Kluwer Academic Publishers, 331–362 (1992)
18. Wu, M., Sakai, H.: getRNIA web software (2013) `http://getrnia.org`
19. Wu, M., Nakata, M., Sakai, H.: An overview of the getRNIA system for nondeterministic data. *Procedia Computer Science* 22, 615–622 (2013)