

プライバシー保護を考慮したアンケートの提案

A Proposal of the Privacy Preserved Questionnaire

劉 臣席 *¹ 中田 典規 *² 酒井 浩 *³
Chenxi Liu Michinori Nakata Hiroshi Sakai

*¹九州工業大学 工学府
Graduate School of Engineering, Kyushu Institute of Technology

*²城西国際大学 経営情報学部
Faculty of Management and Information Sciences, Josai International University

*³九州工業大学 工学研究院
Faculty of Engineering, Kyushu Institute of Technology

In a questionnaire with the selective values, we usually respond one value to each question. Here, we propose such a new framework that we may respond some values to each question. This will dilute the actual value and will be effective to preserve our privacy. *Rough Non-deterministic Information Analysis* (RNIA) is applicable to analyzing such new types of data sets.

1. はじめに

近年、プライバシー保護の必要性が認識され多くの研究が成されている [1, 5, 12]. 我々の研究室では、非決定情報も扱うラフ集合の枠組み、ラフ集合非決定情報解析 (*Rough Non-deterministic Information Analysis:RNIA*) を提案し、不完全性を有する表データからのルール生成を進めている [10, 11, 13]. 今回、RNIA の枠組みを簡単に述べ、これを背景にプライバシー保護を考慮したアンケートを提案する。

2. 研究の背景

2.1 ラフ集合, 非決定情報, DIS と NIS

1980 年代に Pawlak が提案したラフ集合理論 (Rough Set Theory) は、表データから傾向やパターンを読み取るための数学的枠組みであり、表データにおけるデータマイニングや機械学習等へ応用されている [9]. ラフ集合では、通常、確定した属性値から成る決定表 (決定情報システム, DIS とよぶ) の解析を行うが、Pawlak は DIS をあいまいにした非決定情報表 (非決定情報システム, NIS とよぶ) の必要性にも言及している [7, 8]. 例えば、標準的な大学生 A の年齢が不確かでも、日本では「A の年齢は 17 歳以上, 24 歳以下」とすれば真の値がこの区間にあると考えられる. このような情報を非決定情報とよぶ. 図 1 は NIS と起こり得る DIS (派生する DIS とよぶ) の関係を表している. NIS では各属性値は集合で与えられ、その集合に真の値が 1 つ含まれると解釈する. 従って、図 1 は 24 個の派生する DIS の 1 つは真の情報を有するが、どれが真の表であるかは不明であることを表している. NIS は可能性の範囲を考慮することで DIS のあいまいさを補う体系になっており DIS の自然な拡張と考えられる.

連絡先:

劉 臣席 : o100414r@mail.kyutech.jp,
酒井 浩 : sakai@mns.kyutech.ac.jp,
北九州市戸畑区仙水町 1 番 1 号
中田 典規 : nakatam@ieee.org,
東京都千代田区紀尾井町 3 番 2 6 号

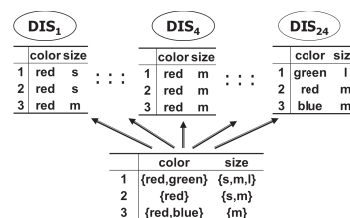


図 1: NIS と 24 個の派生する DIS [11].

NIS については 1980 年代の Lipski による不完全情報データベースの研究 [6] が有名である. この研究では情報検索に確実性と可能性の様相性を導入し、従来にない質問応答を実現している. しかしその後、研究の動向が情報検索からデータマイニングへと進むにつれて、NIS における処理の研究は余り発表されていない. 膨大な場合分け (UCI データレポジトリの Mammographic データセットでは、容易に 10^{100} を超える場合分けが生じる) が必要になるために、多くは何らかの仮定を付けた欠損値 (missing value) の問題として処理されている [3]. しかし、仮定付の欠損値では常に仮定の妥当性の問題が生じ、生成されるルールの意味が不明瞭になる. 我々は、すべての場合分けに依存して様相性を持つルールを定義し、しかもそのルール生成が場合分けの数に依存しない手法を見出した. さらに、その処理系も実現している [11, 13]. 今回、この体系をプライバシー保護の問題に応用する.

2.2 DIS におけるルール

図 1 の DIS を用いてルールを説明する. 属性と属性値の組み、 $[color, red]$ や $[size, s]$ をデスクリプタとよび、デスクリプタを用いた含意式 $\wedge_i [A_i, val_i] \Rightarrow [Dec, val]$ を考える. DIS₁ において含意式 $\tau_1 : [color, red] \Rightarrow [size, s]$ は 2 回生じるので 3 タプルにおける τ_1 の出現率 ($support(\tau_1)$) を $2/3$ とする. また、条件 $[color, red]$ が 3 回起きる中で τ_1 は 2 回起きるので、 τ_1 の無矛盾率 ($accuracy(\tau_1)$) を $2/3$ とする. DIS₄ において、 τ_1 は出現しないので、 $support(\tau_1) = accuracy(\tau_1) = 0$ とする. 含意式 $\tau_2 :$

$[color, red] \Rightarrow [size, m]$ は, $support(\tau_2)=accuracy(\tau_2)=1$ を満たし DIS_4 において τ_2 は最も望ましい含意式になる. 通常, 閾値 α と β に対して $support(\tau) \geq \alpha$ かつ $accuracy(\tau) \geq \beta$ となる含意式 τ を DIS におけるルール (の候補) とする.

2.3 NIS におけるルール

図 1 の NIS において, 非決定情報の定義により派生する DIS の中に真の情報を有する DIS があると考える. そこで, NIS におけるルールを閾値 α と β に対して次のように定義する.

(確実ルール) 派生するどの DIS においても含意式 τ がルールになるとき, この τ を確実ルールとよぶ.

(可能性ルール) 少なくとも 1 つの派生する DIS において含意式 τ がルールになるとき, この τ を可能性ルールとよぶ.

DIS のルールでは非決定情報を考慮していないため, 新たに確実ルールと可能性ルールを導入した [10, 11]. 様相論理の立場において, これらの定義はごく自然なものと考えられる. しかし, 派生する DIS ごとに閾値の判定を行う手法では 10^{100} もある場合分けの逐次処理は難しいと考えられる.

2.4 指標値の最小値と最大値の性質

我々は個別の DIS を列挙する手法ではなく, ラフ集合と粒状計算の手法を用いることにより指標値 $support(\tau)$ と $accuracy(\tau)$ の最小値, 最大値を, 場合分けに依存せず計算できる式を導いた. しかも, $support(\tau)$ と $accuracy(\tau)$ を共に最小にする DIS_{min} の存在と 2 つの値を共に最大にする DIS_{max} の存在を示した [10, 11].

この性質について簡単に言及する. 各含意式 τ はデスクリプタの組み合わせであり, ラフ集合ではデスクリプタ $[A, val]$ を満たす対象集合上の同値類 $eq([A, val])$ を利用する. NIS ではこの同値類を $inf([A, val])$ (値が $\{val\}$ である対象) と $sup([A, val])$ (値に val が含まれる対象) に広げている. 図 1 の DIS_1 において $eq([color, red])=\{1, 2\}$ であり, NIS では $inf([color, red])=\{2\}$, $sup([color, red])=\{1, 2\}$ である. デスクリプタの連言 $\wedge_i[A_i, val_i]$ については以下のようにする.

$$\begin{aligned} inf(\wedge_i[A_i, val_i]) &= \cap_i inf([A_i, val_i]), \\ sup(\wedge_i[A_i, val_i]) &= \cap_i sup([A_i, val_i]). \end{aligned}$$

このように定義した inf と sup では,

$$inf(\wedge_i[A_i, val_i]) \subset eq(\wedge_i[A_i, val_i]) \subset sup(\wedge_i[A_i, val_i])$$

が成立し, NIS Ψ における含意式 $\tau: \wedge_i CON_i \Rightarrow DEC$ に対して以下を得る.

$$\begin{aligned} minsupp(\tau) &= \min_{\phi \in DD(\Psi)} \{ \phi \text{ における } support(\tau) \} \\ &= |inf(\wedge_i CON_i) \cap inf(DEC)| / |OB|, \\ minacc(\tau) &= \min_{\phi \in DD(\Psi)} \{ \phi \text{ における } accuracy(\tau) \} \\ &= \frac{|inf(\wedge_i CON_i) \cap inf(DEC)|}{|inf(\wedge_i CON_i)| + |OUTACC|}, \\ OUTACC &= (sup(\wedge_i CON_i) \setminus inf(\wedge_i CON_i)) \setminus inf(DEC), \\ OB & \text{ は対象の集合, } DD(\Psi) \text{ は派生する } DIS \text{ の集合.} \end{aligned}$$

上記 $minsupp(\tau)$ と $minacc(\tau)$ の定義自体は $DD(\Psi)$ の要素に依存しているが, デスクリプタの inf と sup があれば $DD(\Psi)$ に無関係に 2 式を計算できる. $maxsupp(\tau)$ と $maxacc(\tau)$ も類似した式でかける. $sup(\wedge_i CON_i) \setminus inf(\wedge_i CON_i)$ が変動する部分であり, この変動部分の調整を行うことで上記の式を得ている. DIS_{min} と DIS_{max} の存在もこの点を考慮すれば導き出せる [10, 11].

この結果から, 各 $\phi \in DD(\Psi)$ で計算された点 ($support(\tau)$, $accuracy(\tau)$) を座標平面上にプロットすると図 2 を得る. Mammographic データのように派生する DIS が 10^{100} を超えても全部の点は長方形の内部にプロットされる.

確実ルールの判定は $(minsupp(\tau), minacc(\tau))$ と閾値 α, β

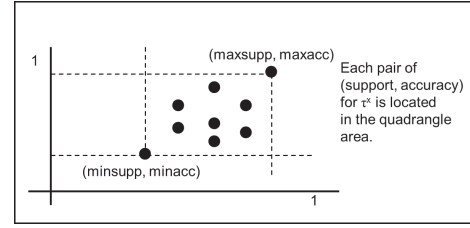


図 2: 指標値の組みのプロット [11].

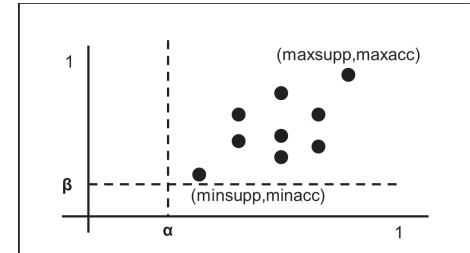


図 3: 含意式 τ が確実ルールであることの判定 [11].

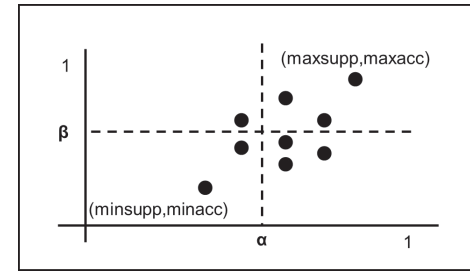


図 4: 含意式 τ が可能性ルールであることの判定 [11].

の比較を行うことで可能であり, この方法は派生する DIS の個数に依存しない (図 3). 同様に, $(maxsupp(\tau), maxacc(\tau))$ と閾値 α, β の比較をすることにより可能性ルールの判定ができる (図 4).

2.5 NIS-Apriori によるルール生成

データマイニングにおける代表的なアルゴリズム, アプリオリ [2] に 2.4 節の性質を応用し, NIS におけるアプリオリ, NIS-アプリオリアルゴリズムを提案した [10, 11]. NIS-アプリオリアルゴリズムは取り出されるルールに対して健全かつ完全である [11]. アプリオリはもともと POS 形式のデータを扱い頻出アイテム集合を取り出す. 我々は, ラフ集合で用いるデスクリプタをアプリオリのアイテムに置き換えることで, 表データを処理する. その際, 確実ルールの生成では $(minsupp(\tau), minacc(\tau))$ と閾値 α, β の比較を行い, 可能性ルールの生成では $(maxsupp(\tau), maxacc(\tau))$ と閾値 α, β の比較を行う. アイテムとしてのデスクリプタの組み合わせの部分は通常のアプリオリと同等であり 2 種類のルールを処理するために, NIS-アプリオリはアプリオリのおよそ倍の計算量が必要と考えられる. しかし, 派生する DIS の個数には依存しない. 最近, NIS-アプリオリを組み込んだツール *getRNIA* をウェブ上に公開した [13]. *getRNIA* は派生する DIS が 10^{100} 以上になる Mammographic データセットから確実ルールや可能性ルールを生成できる.

3. プライバシー保護を考慮したアンケートの提案

上記の研究背景を基にして、複数候補中から質問の回答を1つ選択するアンケートにおいて、下記のアンケート方式を提案する。

新たなアンケートの提案（複数選択回答）

- (1) 回答しにくい内容には1つの選択ではなく複数選択を行えるようにする。つまり、回答者自身で回答に曖昧性を加えることも認める。この回答では *DIS* ではなく *NIS* が得られる。
- (2) RNIA の体系では、データ洗浄で除去されたこのような曖昧性を含むデータも処理対象にできる。元々存在するルールは少なくとも可能性ルールとして取り出すことが保証される。もちろん、確実ルールとして取り出される場合もある。
- (3) 回答の精度を弱めたことにより、アンケートの傾向を確実ルール、可能性ルールにより把握する。

4. データ解析の具体例

ここでは UCI 機械学習レポジトリ中の Lenses データ [4] をアンケートデータと見做し、決定属性 *decision* を特徴付けるルール生成を示し、提案する枠組みを直観的に示す。

4.1 Lenses データ

Lenses データは対象 24 個、5 個の属性 (*age*, *spec*, *astigmatic*, *tear*, *decision*), *age* と *decision* の属性値集合は {1, 2, 3}, その他の属性値集合はいずれも {1, 2} であり、非常に小さいデータである (表 1)。

表 1: Lenses データの一部.

OB	age	spec	astigmatic	tear	decision
1	1	1	1	1	3
2	1	1	1	2	2
3	1	1	2	1	3
4	1	1	2	2	1
5	1	2	1	1	3
:	:	:	:	:	:

4.2 Lenses データにおけるルール生成

我々の研究室では、RNIA in Prolog [10] と *getRNIA* [13] を既に実現しているが、最近、SQL の利用を検討している。*DIS* のデータに対して *apri* と名付けた SQL の *procedure* を用意しており、RDF 形式のデータ (図 5) に対して ‘call *apri*(決定属性, 対象数, サポート, アクキュラシー)’ の実行により 条件部

object	attrib	value
1	age	1
1	spec	1
1	astigmatic	1
1	tear	1
1	decision	3
2	age	1
2	spec	1
2	astigmatic	1

図 5: RDF 形式の Lenses データの一部.

att1	val1	decision	support	accuracy
age	3	3	0.250	0.750 (D1)
tear	1	3	0.500	1.000 (D2)

att1	val1	att2	val2	decision	support	accuracy
astigmatic	1	tear	2	2	0.208	0.833 (D3)
astigmatic	2	spec	2	3	0.208	0.833 (D4)

図 6: Lenses データを特徴付ける 4 個のルール.

が 3 以下の全ルールを得ることができる。RDF 形式はラフ集合におけるデスクリプタそのものの情報であり、*procedure* の中で各表データにおける固有の属性名を属性 *attrib* の属性値として扱うために、同一の *procedure* で任意の表データを処理できる。csv 形式から RDF 形式への変換も容易である。

次に、SQL を利用した処理の例を示す。閾値を *support* \geq 0.2, *accuracy* \geq 0.7 とし、クエリの実行時間 1.2886 秒でルールすべてを含む 11 個の表を生成した。図 6 はルールを保存した 2 つの表である。

4.3 曖昧化された Lenses_NIS データの例

我々は、Lense データに対して意図的に 28 個の欠損値 (*age* に 4 個, *spec* に 7 個, *astigmatic* に 4 個, *tear* に 12 個, *decision* に 1 個) を追加し *NIS* を定義した。紙面において Prolog データ形式の方が多く記載できるので、この形式で次に列挙する。リスト内部のリストの部分が非決定情報である。この小さい *NIS* においても派生する *DIS* の個数は 2038431744 ($=3^{4+1} \times 2^{7+4+12}$) であり、派生する *DIS* の列挙は容易ではないと思われる。

```
data(1, [1, 1, 1, 1, [1, 2, 3]]), data(2, [1, 1, 1, 2, 2]),
data(3, [1, [1, 2], 2, [1, 2], 3]), data(4, [1, 1, 2, 2, 1]),
data(5, [1, 2, 1, [1, 2], 3]), data(6, [1, 2, 1, 2, 2]),
data(7, [1, [1, 2], [1, 2], [1, 2], 3]), data(8, [1, 2, 2, 2, 1]),
data(9, [2, 1, 1, [1, 2], 3]), data(10, [2, 1, 1, 2, 2]),
data(11, [[1, 2, 3], 1, 2, [1, 2], 3]), data(12, [2, 1, 2, 2, 1]),
data(13, [2, [1, 2], 1, [1, 2], 3]), data(14, [2, 2, 1, 2, 2]),
data(15, [2, [1, 2], [1, 2], [1, 2], 3]), data(16, [2, [1, 2], 2, 2, 3]),
data(17, [3, [1, 2], 1, [1, 2], 3]), data(18, [[1, 2, 3], [1, 2], 1, 2, 3]),
data(19, [[1, 2, 3], 1, 2, [1, 2], 3]), data(20, [3, 1, 2, 2, 1]),
data(21, [[1, 2, 3], 2, 1, [1, 2], 3]), data(22, [3, 2, 1, 2, 2]),
data(23, [3, 2, [1, 2], [1, 2], 3]), data(24, [3, 2, [1, 2], [1, 2], 3]).
```

実際に、提案する複数選択回答のアンケートではこのような一部、非決定情報を含む表を得る。この表データを RDF 拡張版形式に変換、同じ閾値を用い、実現している SQL の *procedure* を実行し図 7 の可能性ルールを得た。確実ルールは生成されなかった。つまり、2038431744 個、全部の派生する *DIS* で *support* \geq 0.2 と *accuracy* \geq 0.7 を満たす含意式 τ は存在しない。

Lenses データは Lenses_NIS における派生する *DIS* であり、Lenses におけるルールは少なくとも Lenses_NIS の可能性ルールになる。図 6 と図 7 において、(D1) は (N2), (D2) は (N4), (D3) は (N8), (D4) は (N3) の冗長な場合であり、いずれも可能性ルールとして生成されている。それ以外の (N1), (N5), (N6), (N7) は曖昧性を追加したことにより可能性が広がったために生成されたルールである。(N9) は (N6) の冗長な場合、

att1	val1	deci	deci_value	maxsupp	maxacc	
age	2	decision	3	0.333	0.727	(N1)
age	3	decision	3	0.292	0.778	(N2)
spec	2	decision	3	0.458	0.733	(N3)
tear	1	decision	3	0.542	1.000	(N4)

a1	v1	a2	v2	deci	deci_value	maxsupp	maxacc	
age	1	astigmatic	1	decision	3	0.208	0.714	(N5)
age	1	spec	1	decision	3	0.250	0.750	(N6)
astigmatic	1	spec	1	decision	3	0.292	0.778	(N7)
astigmatic	1	tear	2	decision	2	0.208	0.833	(N8)

a1	v1	a2	v2	a3	v3	deci	deci_value	maxsupp	maxacc	
age	1	spec	1	tear	2	decision	3	0.208	0.714	(N9)
astigmatic	1	spec	1	tear	2	decision	3	0.250	0.750	(N10)

図 7: Lenses_NIS データを特徴付ける 10 個の可能性ルール。

(N10) は (N7) の冗長な場合である。現在の procedure では冗長性の処理が十分にできていない。

従来、アンケートデータにおけるデータマイニングではデータ洗浄の手順によって、処理に不適当なデータは除去されていたと考えられる。今回提案する複数選択回答でも不適当なデータが得られ、Lenses_NIS では 24 タプル中 15 タプルが不適当なデータになると考えられる。しかし、RNIA の体系では 24 タプル全部に依存した様相的な解析が可能になり、従来にない枠組みになっていると考える。各可能性ルールの変動についても、図 2 における 2 点 (*minsupp*, *minacc*) と (*maxsupp*, *maxacc*) の距離を考慮することで対応できると考える。

下記、図 8 は同じ設定において RNIA in Prolog で実行したログファイルである。SQL 版と同じ結果を得ており、Prolog 版と SQL 版共にインプリメントがうまくできていることを保証する例になっていると考える。

5. おわりに

アンケートでは各個人の詳細な情報を得ることが目的でなく、回答者全体としての傾向を知ることを目的としている [5]。この点を考慮すると、回答者自身があいまいな回答をできるに越したことはないと考えられる。今まで、複数選択回答が定着していない理由は、そのアンケートを処理する適当な手法が無かったためと考える。我々は、与えられた非決定情報の下でのデータマイニングを検討してきたが、ラフ集合非決定情報解析では、提案する複数選択回答をそのまま処理できると考える。今回、複数選択回答の提案を行ったが、実際にどの程度有効なアンケートになるかさらなる検討を行いたい。

謝辞：本研究は JSPS 科研費 26330277 の助成を受けたものです。

参考文献

- [1] Aggarwal, C., Yu, S.: Privacy-Preserving Data Mining, Springer, 2008.
- [2] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases, VLDB'94, pp. 487-499, 1994.
- [3] Clark, P., Grzymala-Busse, J.: An analysis of probabilistic approximations for rule induction from incomplete data sets, Fundamenta Informaticae, Vol. 132, No. 3, pp. 365-379, 2014.

```

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
File = [lensesis30207]rs Support= 0.2, Accuracy = 0.7

##### 1st STEP #####
==== Lower System =====
The Rest Candidates: [[[3,1],[5,2]],[[3,1],[5,3]],[[4,2],[5,2]]]
(Next Candidates are Remained)
==== Upper System =====
[6] [age,2] ==> [dec,3] (0.3333333333, 0.7272727273)
Objects:[9,11,13,15,16,18,19,21]
[9] [age,3] ==> [dec,3] (0.2916666667, 0.7777777778)
Objects:[11,17,18,19,21,23,24]
[15] [spec,2] ==> [dec,3] (0.4583333333, 0.7333333333)
Objects:[3,5,7,13,15,16,17,18,21,23,24]
[24] [tear,1] ==> [dec,3] (0.5416666667, 1.0)
Objects:[1,3,5,7,9,11,13,15,17,19,21,23,24]
The Rest Candidates: [[[1,1],[5,3]],[[2,1],[5,3]],[[3,1],[5,3]],[[3,2],[5,3]],[[4,2],[5,2]],[[4,2],[5,3]]]
(Next Candidates are Remained)
EXEC_TIME = 0.0 (sec)

##### 2nd STEP #####
==== Lower System =====
The Rest Candidates: [[[3,1],[4,2],[5,2]]]
(Lower System Terminated)
==== Upper System =====
[1] [age,1]&[spec,1] ==> [dec,3] (0.25, 0.75)
Objects:[1,3,7,11,18,19]
[2] [age,1]&[ast,1] ==> [dec,3] (0.2083333333, 0.7142857143)
Objects:[1,5,7,18,21]
[7] [spec,1]&[ast,1] ==> [dec,3] (0.2916666667, 0.7777777778)
Objects:[1,7,9,13,15,17,18]
[12] [ast,1]&[tear,2] ==> [dec,2] (0.2083333333, 0.8333333333)
Objects:[2,6,10,14,22]
The Rest Candidates: [[[1,1],[4,2],[5,3]],[[2,1],[3,2],[5,3]],[[2,1],[4,2],[5,3]],[[3,1],[4,2],[5,3]],[[3,2],[4,2],[5,3]]]
(Next Candidates are Remained)
EXEC_TIME = 0.0 (sec)

##### 3rd STEP #####
==== Lower System =====
The Rest Candidates: []
(Lower System Terminated)
==== Upper System =====
[1] [age,1]&[spec,1]&[tear,2] ==> [dec,3] (0.2083333333, 0.7142857143)
Objects:[3,7,11,18,19]
[3] [spec,1]&[ast,1]&[tear,2] ==> [dec,3] (0.25, 0.75)
Objects:[7,9,13,15,17,18]
The Rest Candidates: [[[2,1],[3,2],[4,2],[5,3]]]
(Upper System Terminated)
EXEC_TIME = 0.0 (sec)

```

図 8: Lenses_NIS データの RNIA in Prolog による実行。

- [4] Frank, A., Asuncion, A.: UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science, 2010. <http://mllearn.ics.uci.edu/MLRepository.html>
- [5] 菊池浩明: データマイニングと個人情報保護, 第 3 回情報科学技術フォーラム, FIT2004 講演論文集, 2004.
- [6] Lipski, W.: On databases with incomplete information, Journal of the ACM, 28(1), pp. 41-70, 1981.
- [7] Orłowska, E., Pawlak, Z.: Representation of nondeterministic information, Theoretical Computer Science, 29(1-2), pp. 27-39, 1984.
- [8] Pawlak, Z.: Systemy informacyjne: Podstawy teoretyczne (in Polish), Wydawnictwo Naukowo-Techniczne, Poland, 1983.
- [9] Pawlak, Z.: Rough Sets, Kluwer Academic Publishers, 1991.
- [10] Sakai, H., Okuma, H., Wu, M., Nakata, M.: Rough non-deterministic information analysis for uncertain information, The Handbook on Reasoning-Based Intelligent Systems, World Scientific, pp. 81-118, 2013. <http://www.mns.kyutech.ac.jp/sakai/RNIA>
- [11] Sakai, H., Wu, M., Nakata, M.: Apriori-based rule generation in incomplete information databases and non-deterministic information systems, Fundamenta Informaticae, Vol. 130, No. 3, pp. 343-376, 2014.
- [12] 佐久間淳, 小林重信: プライバシー保護データマイニング, 人工知能学会誌, 第 24 巻第 2 号, 2009.
- [13] Wu, M., Nakata, M., Sakai, H.: An overview of the getRNIA system for non-deterministic data, Procedia Computer Science, Elsevier, Vol. 22, pp. 615-622, 2013. <http://getrnia.org/>