ラフ集合非決定情報解析とNIS-アプリアリシステム —可能世界意味論に基づくルール生成システム—†

酒井 浩*¹・中田 典規*²・和多田 淳三*³

この論文では、表データにおける情報の不完全性と表データにおけるルール生成に注目し、今までに提案された重要な研究をラフ集合とデータマイニングの観点から解説する。また、著者らが提案している「ラフ集合非決定情報解析」と「NIS-アプリアリシステム」について述べ、不完全な情報表に基づくルール生成を概説する。ラフ集合非決定情報解析は可能世界意味論に基づくラフ集合を提案し、情報の不完全性を考慮したルール生成問題に一解法を与えている。この解法アルゴリズムはNIS-アプリアリアルゴリズムと名付けられ、NIS-アプリアリシステムとしてSQL言語により実装されている。NIS-アプリアリアルゴリズムをコアアルゴリズムにするラフ集合非決定情報解析の新たな可能性にも言及する。

キーワード：ルール生成、ラフ集合、非決定情報、可能世界意味論、アプリアリアルゴリズム

1. はじめに

ラフ集合による離散値表データ解析の出発点はMarekとPawlakによる関係データベースの数理的手法[1]にあると考える。この当時は、関係代数、情報検索や問い合わせ処理の研究[2]が進められており、Lipskiの不完全情報データベース[3]やOrłowskaとPawlakによる非決定情報システム（Nondeterministic Information System, NIS）[4]が提案されている。これらは、情報の不完全性に関する重要な枠組みとして知られており、実際、著者らの研究はNISの枠組みを継承している。情報の不完全性をメンバシップ値等の数値ではなく、場合分けによる様相性で処理する点にこれらの研究の特徴がある。

1980年頃になると表データの情報検索や問い合わせ処理に加えて表データからの知識獲得やデータマイニングの研究も始まり、1982年にPawlakによるラフ集合[5]が提案されている。Pawlakは表データで自然に定義できる類の包含関係により無矛盾な含意式の判定が可能になることを利用し、無矛盾な含意式の生成を行っている[5,6]。無矛盾な含意式の生成から出発したルール生成の研究は、その後、多様な体系に発展している[7–13]。本学会誌でも種々の解説が成されている[14–19]。

当然、不確実な表データにおけるルール生成問題も研究されており、Grzymała-BusseによるLERSシステム[20,21]やKryszkiewiczによる不完全性ルール[22]の研究は良く知られ

ている。ただ、これらの研究はNISの枠組みではなく欠損値（missing value）を含む表データ[20,22]におけるルール生成を議論している。これらの研究では欠損値に対する特定の解釈が定められ、欠損値を考慮した同値類の演算によりルールが生成される。著者らの枠組みはこれらの研究に類似しているが、著者らの可能世界意味論によるルールの定義と前者のルールの定義は異なる（参考文献[23]にルールの差異を示す具体例がある）。著者らはNISと可能世界意味論の枠組みを陽に採用し、様相性を含むルール生成を進めている。確実性と可能性を理論的に定義できる可能世界意味論の枠組みは[16,17]においても解説されている。

しかし、実際のルール生成においては可能世界数が指数関数的に増加する問題がある。UCIデータレポジトリにあるMammographic MassデータやCongressional Votingデータ[24]では可能世界数が 10^{100} を超える。可能世界処理への対応が進まない中で、NISの枠組みは欠損値の枠組みに置き換わったと思われる。この後、NISの枠組みに基づく研究は殆ど見られないが、数年前に著者らは多数の可能世界において特徴的な2可能世界を調べることでルール生成が可能になることを証明した[23]。これにより、NISにおけるルール生成が初めて現実のものになったと考える[23,25–27]。

本論文では過去のルール生成の研究を列挙し、著者らが行った下記についても述べる。

- (1) 決定情報表のラフ集合データ解析を非決定情報表のラフ集合に広げたデータ解析の体系（これをラフ集合非決定情報解析、Rough sets Non-deterministic Information Analysis：RNIA[25]とよんでいる）、
- (2) RNIAで定義した確実ルール、可能性ルール生成のアルゴリズム（これをNIS-アプリアリアルゴリズム、NIS-Apriori algorithm[23,26,28]とよんでいる）、
- (3) SQLによるルール生成システムの実現[26,29]。

本論文の構成を次に示す。第2節で決定情報表におけるラフ集合とルール生成について概説する。第3節ではトランザクションデータを処理するためのアプリアリアルゴリズム[30]

† Rough Sets Non-Deterministic Information Analysis and a NIS-Apriori System – A Rule Generation System Based on Possible World Semantics –
Hiroshi SAKAI, Michinori NAKATA, and Junzo WATADA

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

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

*3 マレーシア、ペトロナス大学 情報学部
Department of Computer & Information Sciences, Universiti Teknologi PETRONAS, Malaysia

表 1 DIS の例 ψ

OB	$size$	$weight$	$price$
1	<i>small</i>	<i>light</i>	<i>low</i>
2	<i>small</i>	<i>heavy</i>	<i>low</i>
3	<i>large</i>	<i>normal</i>	<i>high</i>
4	<i>large</i>	<i>light</i>	<i>high</i>
5	<i>large</i>	<i>light</i>	<i>low</i>

を決定情報表のアプリオリアルゴリズムに転換する. 第 4 節では NIS におけるラフ集合を提案し, ルール生成のための重要な定理を示す. また, SQL 言語による NIS-アプリオリアルゴリズムの実行例を示す. 第 5 節において本論文をまとめ, RNIA の新たな可能性にも言及する. RNIA は離散値の手続きを行うため, ルール生成を行う過程の副産物として真の情報の推定が可能になり, 情報の不完全性の側面からプライバシー保護への応用も考えられる.

2. 決定情報表におけるラフ集合とルール生成

NIS と区別するために通常の情報表を決定情報表 (Deterministic Information System, DIS) とよぶことにする. また, 詳細な定義は [14–19] に従う.

2.1 基本的な枠組み

DIS を 4 つ組 $\psi = (OB, AT, VAL_A (A \in AT), f)$ (OB : 有限集合で各要素を対象とよぶ, AT : 有限集合で各要素を属性とよぶ, VAL_A : 有限集合で各要素を属性値とよぶ, 写像 $f: OB \times AT \rightarrow \cup_A VAL_A$) で定義する. 属性 A と属性値 $val_A \in VAL_A$ の組 $[A, val_A]$ をデスクリプタとよぶ. デスクリプタは DIS ψ の特徴を記述するための単語の役割を果たし, その設定は重要である. ψ では AT も VAL_A も有限集合であるので自然に有限個のデスクリプタを得るが, VAL_A が実数値の区間などの場合には何らかの離散化によりデスクリプタを定義する必要がある [31]. 通常, AT には 1 つの決定属性 Dec が含まれているとし, $AT \setminus \{Dec\}$ の部分集合 CON を条件属性の集合とよぶ. 対象 x で定義される式 $\tau_x: \wedge_{A \in CON} [A, f(x, A)] \Rightarrow [Dec, f(x, Dec)]$ を x の含意式とよび, 優れた特徴をもつ含意式 τ_x をルールとよぶ. 優れた特徴を定義する指標として $support(\tau)$, $accuracy(\tau)$ と $coverage(\tau)$ を示しておく. 3 指標はラフ集合 [6, 10, 11] にも Agrawal のアプリオリアルゴリズム [30] にも用いられている.

定義 1 [6, 10, 11, 30]

- (1) $support(\tau) = |eq(\wedge_i DES_i) \cap eq([Dec, v])| / |OB|$,
 - (2) $accuracy(\tau) = |eq(\wedge_i DES_i) \cap eq([Dec, v])| / |eq(\wedge_i DES_i)|$,
 - (3) $coverage(\tau) = |eq(\wedge_i DES_i) \cap eq([Dec, v])| / |eq([Dec, v])|$.
- ただし, 一般の含意式 $\tau: \wedge_i DES_i \Rightarrow [Dec, v]$ (DES_i はデスクリプタ) に対して, $eq([A_i, val_i]) = \{x \in OB \mid f(x, A_i) = val_i\}$, $eq(\wedge_i DES_i) = \cap_i eq(DES_i)$, $|*|$ 記号は集合 $*$ の要素数を表す. また, $eq()$ をデスクリプタで決まる OB 上の同値類という.

例 1 表 1 の DIS ψ を用いて種々の定義を例で示す.

(1) DIS ψ では, $OB = \{1, 2, 3, 4, 5\}$, $AT = \{size, weight, price\}$, $Dec = price$, CON は $\{size, weight\}$ の部分集合である.

(2) 含意式 $\tau: [size, large] \Rightarrow [price, high]$ において $eq([size, large]) = \{3, 4, 5\}$, $eq([price, high]) = \{3, 4\}$ なので下記を得る.

$$support(\tau) = \frac{|eq([size, large]) \cap eq([price, high])|}{|OB|} = \frac{|eq([price, high])|}{5} = \frac{2}{5},$$

$$accuracy(\tau) = \frac{|eq([size, large]) \cap eq([price, high])|}{|eq([size, large])|} = \frac{2}{3},$$

$$coverage(\tau) = \frac{|eq([size, large]) \cap eq([price, high])|}{|eq([price, high])|} = \frac{2}{2} = 1.0.$$

$support(\tau)$ は τ の出現の割合, $accuracy(\tau)$ は τ の無矛盾性の割合, $coverage(\tau)$ は τ による結論部の被覆の割合を表す. 各指標値とも 0 以上 1 以下の値をとり, 1 に近い値を取る方が優れた特徴を有している.

(3) 含意式 $\tau': [size, large] \wedge [weight, normal] \Rightarrow [price, high]$ は τ の条件部に別のデスクリプタが追加されており, 一般に $support(\tau') \leq support(\tau)$ が成立する. 一方, $accuracy$ 値については, $accuracy(\tau') = 1/1 = 1.0 \geq accuracy(\tau)$ であるが, 別の含意式 $\tau'': [size, large] \wedge [weight, light] \Rightarrow [price, high]$ において $accuracy(\tau'') = 1/2 = 0.5 \leq accuracy(\tau)$ である. つまり, 条件部が追加された含意式の $accuracy$ 値は元の含意式の $accuracy$ 値と比較して増加することも減少することもある.

2.2 縮約計算によるルール生成

本節では, Pawlak の方法 [5, 6] (縮約法とよぶ) に言及する. Pawlak の研究において次の命題は重要な役割を果たしている.

命題 1 [6] DIS における含意式 $\tau: \wedge_i DES_i \Rightarrow [Dec, v]$ において下記 (1) と (2) は同値である.

- (1) $accuracy(\tau) = 1$ (このとき τ は無矛盾であるという),
- (2) $eq(\wedge_i DES_i) \subseteq eq([Dec, v])$.

命題 1 より, 無矛盾な含意式を見出すためには, 結論部の条件で決まる同値類に対してその部分集合になる同値類を定義するデスクリプタを見出せばよいことになる. 例 1 において $eq([price, high]) = \{3, 4\}$ なので, $\{3, 4\}$ の部分集合を定義するデスクリプタを条件部に指定すれば無矛盾な含意式を得る. $eq([size, large]) = \{3, 4, 5\}$ では包含関係は成立しないが, $eq([size, large] \wedge [weight, normal]) = \{3\}$ では包含関係が成立し無矛盾な τ' を得る. この例では条件部が 1 デスクリプタでは不都合であり, 条件部を 2 デスクリプタに増やしている.

しかし, 実際の DIS における τ の選択について命題 1 は何も言及していない. Pawlak はさらに縮約の概念を導入している [6, 15]. 一般に, 属性集合 P で定義される同値類の集合と P の真部分集合 Q で定義される同値類の集合が一致し, かつ Q のいかなる真部分集合で定義される同値類の集合とも一致しない場合, Q を P の縮約という. また, 決定属性が定義する同値類の正領域を保持する場合の縮約を特に相対縮約 [6] とよぶ. 縮約, 相対縮約とも一意に決まることは稀であり, 一般に複数

存在する。相対縮約は結論部を強く支持するための条件属性集合を意味する。相対縮約の計算には、上記の例のように属性を追加する場合と、逆に全条件属性を用いた含意式（無矛盾になる場合もならない場合もある）から出発し、無矛盾であると確認された各含意式の無矛盾性を維持しながら、できる限り多くの条件属性を除去する場合がある。1つの相対縮約が与えられれば、DISの各対象に対して命題1を応用できる。

また、ルール τ の条件部からどのデスクリプタを除去しても矛盾になる場合、この τ は極小[6, 10, 20]であると定義され、すべてのルールではなく、すべての極小ルールの生成が考えられている。極小ルールの条件部にデスクリプタが追加された含意式は冗長であると見なされ扱われていない。

注意 1 Pawlak の縮約法は 2 つの要件、 $accuracy(\tau) = 1$ (*support* の要件は用いられていない) と極小である、をみたすすべての含意式を生成するための一手法と見ることができ。

2.3 識別関数法によるルール生成

本節では、Skowron の方法 [10]（識別関数法とよぶ）に言及する（この手法については [15] にも詳しい解説がある）。Skowron は Pawlak の縮約計算を発展させ 2 つの要件、 $accuracy(\tau) = 1$ と極小である、をみたすすべての含意式を生成する別の手法を提案している。例 1 において、 $[price, low]$ に関する場合を考える。 $[price, low]$ によって同値類 $\{1, 2, 5\}$ （ターゲット集合とよぶ）が決まるので、ターゲット集合以外の対象を対象 1, 2, 5 から区別できればよい。対象 1 と対象 3 を区別する条件は $[size, small] \vee [weight, light]$ 、対象 1 と対象 4 を区別する条件は $[size, small]$ 、対象 2 と対象 3 を区別する条件は $[size, small] \vee [weight, heavy]$ 、対象 2 と対象 4 を区別する条件は $[size, small] \vee [weight, heavy]$ 、対象 5 と対象 3 を区別する条件は $[weight, light]$ 、対象 5 と対象 4 を区別する条件はない。従って、対象集合 $\{1, 2\}$ に関する識別関数 $([size, small] \vee [weight, light]) \wedge ([size, small]) \wedge ([size, small] \vee [weight, heavy])$ を得る。これを論理関数と見た場合のデスクリプタの解集合を $[price, low]$ の条件部にすることで無矛盾な含意式を得ることができ、極小な含意式 $[size, small] \Rightarrow [price, low]$ を得る。別の解集合に対応する $[size, small] \wedge [weight, light] \Rightarrow [price, low]$ や $[size, small] \wedge [weight, heavy] \Rightarrow [price, low]$ も $accuracy(\tau)=1$ をみたす含意式になる。

Skowron はターゲット集合を識別するための論理関数を作り、その解の連言を条件部に置いた τ は $accuracy(\tau) = 1$ をみたすことを示している。従って、論理関数の極小な解を全部列挙することで、 $accuracy(\tau) = 1$ をみたすすべての極小な含意式を得ることができる。具体的ルール生成の例は [32] にある。一方、すべての相対縮約を得る問題は NP 困難な問題になることがグラフの頂点被覆問題に基づいて示されている [10, 33]。論理関数のすべての極小な解を得る問題が制約充足問題 SAT であることから、 $accuracy(\tau) = 1$ をみたすすべての極小な含意式を得る問題の NP 困難性を把握できる。

注意 2 Skowron の結果から 2 つの要件、 $accuracy(\tau) = 1$ と極小である、をみたすすべての含意式を得る問題は表データのサイズが大きくなると難しくなる。このためか、過去の研究ではすべてのルールを取り出す（見落とさなくすべてのルールを生成する、アルゴリズムの完全性として扱われる）議論はあまり見当たらない。得られたアルゴリズムの妥当性を k -分割交差検証により確認する研究が多い。しかし、アルゴリズムに完全性の保証がなければ重要なルールを見落とす危険性がある。

2.4 被覆計算によるルール生成

本節では、被覆計算によるルール生成に言及する。ルール生成では、先に示した縮約法、識別関数法や次に示す Agrawal のアプリアリ法のように含意式がみたす要件を指定する枠組み以外に、得られる含意式の集合がみたす要件を指定する場合もある。決定木 [34] では分類のエントロピーを小さくする手順で得られた含意式（決定木の枝を条件部、葉を結論部と見る）をルールとし、支配に基づくラフ集合 [7] では決定属性の順序を保存する手順で得られた含意式をルールとしている。一般に、被覆系のアルゴリズム [20] はターゲット集合を説明するルールの集合を生成し、ターゲット集合のどの対象もいずれかのルールで結論付けられる、つまりルールの集合として $coverage = 1$ をみたすルール生成であると考えられる。しかし、個々の含意式に対する *support* 値や *accuracy* 値は後付けで得られるため、*support* 値や *accuracy* 値をルール生成の要件にする枠組みとは異なる。また、ターゲット集合を被覆する複数の類がある場合、類の選択の順序により異なるルールが得られると考えられる。本論文では、*support* 値と *accuracy* 値でルールを定義する枠組みを扱うために、被覆計算によるルール生成には触れない。詳細な解説は [19] にある。

2.5 得られたルールとルール生成法の妥当性

ルール生成の研究には次の 3 研究体系があると考えられる。

- (1) ルールの定義を与え、そのルールを生成する手法を示す研究（Pawlak の縮約法、Skowron の識別関数法など）、
- (2) 何らかの要件をみたす操作を行い、結果としてルールの集合を得る研究（決定木、被覆系のアルゴリズムなど）、
- (3) 主に、近似によるルールの定義を追及する研究。

(1) の Pawlak や Skowron の枠組みでは $accuracy(\tau) = 1$ をみたす τ はルールとよばれ、先に示した縮約法や識別関数法は $accuracy(\tau) = 1$ をみたす極小な τ を過不足なくすべて見出すことができる。一般に、この「過不足なく」の性質を有する手法は論理的に健全かつ完全な手法であるとよばれている。健全性・完全性の性質は論理に基づく計算の妥当性を保証する重要な概念である。例えば、論理プログラムにおける導出計算の妥当性 [35] や Lipski の不完全情報データベースにおける公理系の妥当性 [3] は本概念により理論的に保証されている。同様に、ルール集合に対する健全性・完全性は実験によらず理論的に生成アルゴリズムの妥当性を保証する一要因になる。(2) の体系では個々のルール自体の定義が成されていないため、ルール集合に対する健全性や完全性の概念はなじまない。そ

表2 トランザクションデータ（買い物）の例

買い物番号	アイテム集合
1	{bread, milk, ham, beer, corn}
2	{cheese, ham, beer}
3	{ham, beer, apple, potato, corn}
4	{cheese, cake, beer}

して、(2)の体系では課する要件の性質や k -分割交差検証によりルールと生成アルゴリズムの妥当性を保証する場合が多い。(3)の体系では、Yao らにより decision-theoretic rough sets や multi-granulation rough sets が提案されている [9, 12].

3. アプリオリ法とルール生成

本節では Agrawal によるもう 1 つのルール生成法（アプリオリ法とよぶ）に言及する。

3.1 基本的な枠組み

アプリオリ法では、例えば表2のようなデータを扱う。品物を一般にアイテムとよび、買い物番号に品物集合（アイテム集合とよぶ）を対応付けた表である。

表2においてアイテム集合 {ham, beer} に注目すると、このアイテム集合は4買い物中3回起きており、ham を買った人は必ず beer を買っている。 $\tau : \text{ham} \Rightarrow \text{beer}$ を含意式と見ると指標値 $\text{support}(\tau) = 3/4$ 、 $\text{accuracy}(\tau) = 3/3 = 1.0$ である。このようにしてアイテムを項とする含意式に2つの指標値を対応させるルール生成を行う。

定義2 [30]（Agrawal のルール生成問題）

トランザクションデータにおいて、指標値 α と β ($0 < \alpha, \beta \leq 1.0$) を定める。このとき、 $\text{support}(\tau) \geq \alpha$ かつ $\text{accuracy}(\tau) \geq \beta$ をみたすすべての含意式 τ を生成せよ。

定義2の問題を解決する手法がアプリオリ法である。アプリオリ法はルールの候補になる含意式を逐次生成・判定し、全含意式の指標値を調べることと同等になる手続きを実現している。

例2 表2において support 値が 0.5 以上になるアイテム集合（ support 値の要件をみたすアイテム集合を頻出アイテム集合とよぶ）の生成過程を考える。対象数が4であり $\text{FREQ}/4 \geq 0.5$ より $\text{FREQ} \geq 2$ 、つまりアイテム集合は2回以上出現する必要がある。要素数1のアイテム集合の（頻度付きの）候補集合は

$$C_1 = \{\{\text{bread}\}(1), \{\text{milk}\}(1), \{\text{ham}\}(3), \{\text{beer}\}(4), \{\text{corn}\}(2), \{\text{cheese}\}(2), \{\text{apple}\}(1), \{\text{potato}\}(1), \{\text{cake}\}(1)\}$$

であり、頻出アイテム集合の集合は

$$I_1 = \{\{\text{ham}\}(3), \{\text{beer}\}(4), \{\text{corn}\}(2), \{\text{cheese}\}(2)\}$$

になる。 I_1 を使い、要素数2の頻出アイテム集合の候補集合を作る。 $C_2 = \{P \cup Q \mid P, Q \in I_1 (P \neq Q)\}$ とし、頻度を確認する

表3 DIS をトランザクションデータと見なす例

OB	アイテム集合
1	{[size,small], [weight,light], [price,low]}
2	{[size,small], [weight,heavy], [price,low]}
3	{[size,large], [weight,normal], [price,high]}
4	{[size,large], [weight,light], [price,high]}
5	{[size,large], [weight,light], [price,low]}

ことで

$$I_2 = \{\{\text{ham, beer}\}(3), \{\text{ham, corn}\}(2), \{\text{beer, corn}\}(2), \{\text{beer, cheese}\}(2)\}$$

を得る。さらに、要素数3の頻出アイテム集合の候補を I_2 の要素から作る。例えば、{ham, corn} と {beer, cheese} から {ham, corn, beer}, {ham, corn, cheese}, {beer, cheese, ham} と {beer, cheese, corn} を候補にする。候補の頻度を確認し、 $I_3 = \{\{\text{ham, beer, corn}\}(2)\}$ を得る。これより、 support 値が 0.5 以上になるすべての頻出アイテム集合の集合 I_1, I_2 と I_3 を得ることができる。アプリオリ法は頻出アイテム集合の集合を要素数の順に列挙し、その後、決定部を決めた含意式に対する accuracy 値の判定を行い、すべてのルールを得ている [30].

定義2と類似する研究に Ziarko の VPRS (Variable Precision Rough Sets) [13] がある。ターゲット集合 $X (eq([Dec, v]) = X)$ と条件部（デスクリプタの連言 $\wedge_i [A_i, val_i]$ ）が定義する類 E から誤判別率 $C(E, X)$ を計算し、 $C(E, X) \leq \beta$ ($0 \leq \beta < 0.5$ は閾値) により含意式 $\wedge_i [A_i, val_i] \Rightarrow [Dec, v]$ の妥当性を判定する。VPRS により、ルールの要件が $\text{accuracy} = 1$ から $\text{accuracy} \geq 1 - \beta$ へと広げられている。2手法とも類似した枠組みであるが、ルール生成の実行面で比較すると、VPRS は誤判別率、つまり accuracy 値を優先にした探索、アプリオリ法は support 値を優先にした探索になっている。すべてのルールを得る立場から見ると、 support 値の単調性を利用できるアプリオリ法が効率的かと考えられる。

3.2 決定情報表におけるアプリオリ法

表3は表1をトランザクションデータに変換した例である。表1をデスクリプタの集合と見て、各対象ごとのデスクリプタの集合を記述している。この変換により、表データにアプリオリ法の手法を持ち込む。

決定情報表におけるアプリオリ法では次の特徴があり、これらを考慮しアプリオリ法の手順を真似しながら DIS のアプリオリ法を提案する。

- (1) 決定属性が決まっており、元データにおけるアイテム集合の要素数は属性数と一致する。
- (2) ルール生成時に列挙するアイテム集合は決定属性のデスクリプタを1個だけ含み、異なる条件属性のデスクリプタを複数個含むものに限定する（この集合から自動的に1つの含意式が決まる）。例2において、アプリオリ法では I_1, I_2, I_3 を得た後 accuracy 値の判定を行うが、DIS のアプリオリ法では決定部が決まっているために accuracy 値の判定も行いながら頻出

Require: DIS ψ , 決定属性 Dec , 閾値 α, β .

Ensure: 極小ルールの集合 $Rule = \cup_{i=1} Rule_i$.

```

1:  $i \leftarrow 1$ ;
2:  $SubIMP_i = \{\tau_{i,j} \in IMP_i \mid support(\tau_{i,j}) \geq \alpha\}$  を作る;
3: while ( $|SubIMP_i| \geq 1$ ) do
4:    $Rest_i \leftarrow \{\}; Rule_i \leftarrow \{\};$ 
5:   for all  $\tau_{i,j} \in SubIMP_i$  do
6:     if  $accuracy(\tau_{i,j}) \geq \beta$  then  $\tau_{i,j}$  を  $Rule_i$  に追加;
7:     else  $\tau_{i,j}$  を  $Rest_i$  に追加;
8:     end if
9:   end for
10:   $i \leftarrow i + 1$ ;
11:   $Rest_{i-1}$  を用い  $SubIMP_i = \{\tau_{i,j} \in IMP_i \mid support(\tau_{i,j}) \geq \alpha\}$ 
12:  を作る;
13: end while
14: return  $Rule = \cup_{i=1} Rule_i$ 

```

図 1 DIS-Apriori アルゴリズム

アイテム集合を列挙する.

ルール生成・利用の立場では縮約法で用いられた極小の概念は自然であり, 可能な限り条件部が小さいルールが望ましい. そこで著者らは Pawlak の極小の概念に従い, 次の極小ルールを扱う (定義 2 の Agrawal のルール生成ではこの極小の概念は用いられていない).

定義 3 DIS ψ において, 決定属性を Dec , 指標値 α と β ($0 < \alpha, \beta \leq 1.0$) を定める. $support(\tau) \geq \alpha$ かつ $accuracy(\tau) \geq \beta$ をみたす $\tau: \wedge_i DES_i \Rightarrow [Dec, v]$ において, $\tau': \wedge_{DES_j \in M} DES_j \Rightarrow [Dec, v]$ (M は空でない $\{DES_i\}$ の任意の真部分集合) がルールにならない場合, τ を極小ルールとよぶ.

定義 4 (DIS におけるルール生成問題)

DIS ψ において, 決定属性を Dec , 指標値 α と β ($0 < \alpha, \beta \leq 1.0$) を定める. $support(\tau) \geq \alpha$ かつ $accuracy(\tau) \geq \beta$ をみたすすべての極小ルール τ を生成せよ.

まず単純な手法を示す. $IMP_1 = \{\tau \mid \tau: [A, val_A] \Rightarrow [Dec, v]\}$ (条件部が 1 デスクリプタの全含意式の集合), IMP_2 (条件部が 2 デスクリプタの全含意式の集合), IMP_3 (条件部が 3 デスクリプタの全含意式の集合), ..., IMP_n (n は条件属性数) とする. 全含意式は $\cup_{i=1}^n IMP_i$ であり, 各含意式を逐次調べることでルール生成問題を解決できる. しかし, この単純な手法の利用には実行時間上の難点がある. DIS ψ において, $|OB| = 1000$, $|AT| = |CON \cup \{Dec\}| = 11$ の場合, 各対象 x について ${}_{10}C_1$ (IMP_1 の要素数), ${}_{10}C_2$ (IMP_2 の要素数), ..., ${}_{10}C_{10}$ (IMP_{10} の要素数) 個の含意式がある. 二項定理の性質から ${}_{10}C_1 + {}_{10}C_2 + \dots + {}_{10}C_{10} = (1+1)^{10} - 1 = 1024 - 1 = 1023$ となり, ψ において調べる含意式は $1023 \times 1000 \approx 10^6$ 個になる. 10^6 個の含意式を列挙して逐次指標値を計算する手法には無理があると感じられる.

著者らは, 図 1 の DIS-アプリオリ (アルゴリズム) を提案する. DIS-アプリオリでは, 1 回目の while ループで IMP_1 に

```

mysql> show tables;
+-----+
| Tables_in_soft |
+-----+
| rdf            |
| table 1        |
+-----+
2 rows in set (0.00 sec)

mysql> call apriv3('price',5,0.1,0.7);
Query OK, 0 rows affected (0.85 sec)

mysql> show tables;
+-----+
| Tables_in_soft |
+-----+
| con1           |
| con2           |
| con3           |
| condi          |
| deci           |
| rdf            |
| rest1          |
| rest2          |
| rest3          |
| rule1          |
| rule2          |
| rule3          |
| table 1        |
+-----+
13 rows in set (0.00 sec)

```

図 2 apriv3 プログラムによる表 1 データからの全極小ルール ($support(\tau) \geq 0.1$, $accuracy(\tau) \geq 0.7$) の生成. それぞれの極小ルールは各テーブルに登録されている

含まれる極小ルールを $Rule_1$ に追加, 2 回目の while ループで IMP_2 に含まれる極小ルールを $Rule_2$ に追加, ..., k 回目の while ループで IMP_k に含まれる極小ルールを $Rule_k$ に追加する. 条件属性数は有限なので DIS-アプリオリは必ず停止する. ここまでは上記の単純な手法と同じであるが, 一部の含意式 $SubIMP_i$ ($\subseteq IMP_i$) のみを調べ, IMP_i を調べたことと同等の効果を得ることができるような工夫がされている.

3.3 SQL による DIS-アプリオリシステム

表 1 の実行例を通して DIS-アプリオリ自体と処理システムの動きを説明する. 図 2 では, $Dec = price$, 対象数 5, $support(\tau) \geq 0.1$, $accuracy(\tau) \geq 0.7$ の要件をみたすすべての極小ルール τ を生成し, 結果をテーブル $Rule_1$ (条件部 1 個), $Rule_2$ (条件部 2 個), $Rule_3$ (条件部 3 個) に書き込んでいる. $support(\tau) \geq 0.1$ より $|OB| \times \alpha = 5 \times 0.1 = 0.5$ となり τ を構成するデスクリプタは 1 回以上生じればよい. この例では, 各デスクリプタは候補になる (図 3). 仮に, $\alpha = 0.3$ とすると 2 回以上出現するデスクリプタのみ扱い, $[weight, heavy]$ や $[weight, normal]$ を含む含意式はこの時点で除外される. この例では, $5 \times 2 = 10$ 個の含意式を調べ, 下記を行う (図 3).

(1) $support(\tau) \geq 0.1$ と $accuracy(\tau) \geq 0.7$ をみたす 3 含意式を $Rule_1$ に追加し, 3 含意式を判定済とする. τ の冗長な含意式 τ' は極小ルールにはならないため, τ' も判定済とする.

(2) $support(\tau) \geq 0.1$ と $accuracy(\tau) < 0.7$ をみたす 4 含意式を $Rest_1$ に追加し, 4 含意式を未判定とする. 例 1 で示したように未判定な含意式の条件部に適当なデスクリプタを追加することにより, ルールを得る場合がある.

(3) $support(\tau) < 0.1$ をみたす 3 含意式はルールにならないため, 3 含意式を判定済とする. τ の冗長な含意式 τ' は $support(\tau') \leq support(\tau) < \alpha$ をみたすため, τ' はルールにな

```
mysql> select * from con1;
```

att1	val1	freq
size	large	3,000
size	small	2,000
weight	heavy	1,000
weight	light	3,000
weight	normal	1,000

5 rows in set (0.00 sec)

```
mysql> select * from deci;
```

deci	deci_value	freq
price	high	2,000
price	low	3,000

2 rows in set (0.00 sec)

```
mysql> select * from rule1;
```

att1	val1	deci	deci_value	support	accuracy
size	small	price	low	0.400	1.000
weight	heavy	price	low	0.200	1.000
weight	normal	price	high	0.200	1.000
end_attrib	NULL	NULL	NULL	NULL	NULL

4 rows in set (0.00 sec)

```
mysql> select * from rest1;
```

att1	val1	deci	deci_value	support	accuracy
size	large	price	high	0.400	0.667
size	large	price	low	0.200	0.333
weight	light	price	high	0.200	0.333
weight	light	price	low	0.400	0.667

4 rows in set (0.00 sec)

図3 得られた極小ルール, $Rest_1$ の内容

らない。従って, τ' も判定済とする。

DIS-アプリアリでは $support$ 値をみたし未判定の含意式の候補集合 $SubIMP_i$ ($i \geq 2$) を作る。ここでは、頻出アイテム集合の要件以外に未判定の要件も追加されており、DIS-アプリアリによる候補集合はアプリアリ法による候補集合をさらに絞っている。

命題2 $SubIMP_2 = \{p \wedge q \Rightarrow r \mid support(p \wedge q \Rightarrow r) \geq \alpha, p \Rightarrow r, q \Rightarrow r \in Rest_1\}$ とかける。図1の $Rest_1$ における同じ結論部を持つ含意式の組み合わせを行い、 $support$ 値の判定を行えば $SubIMP_2$ を得る。

(証明) 含意式 $p \Rightarrow r \in IMP_1$ の条件部にデスクリプタ q を追加し、条件部が2個のデスクリプタである含意式 $p \wedge q \Rightarrow r \in IMP_2$ を作り $SubIMP_2$ を決める。含意式 $p \Rightarrow r$ に対して、デスクリプタ q を含む $q \Rightarrow r$ と $q \Rightarrow s$ ($s \neq r$) を考えればすべての場合を考慮したことになる。このとき、含意式 $p \Rightarrow r$ と $q \Rightarrow s$ は結論部が異なるため $p \Rightarrow r$ をみたす対象は $q \Rightarrow s$ を満たさない。つまり、 $p \wedge q \Rightarrow r$ は表に出現しない(しかし $q \Rightarrow s$ と独立に $q \Rightarrow r$ をみたす対象があれば表に出現することもある)。そこで、含意式 $p \Rightarrow r$ と $q \Rightarrow r$ の関係を考える。含意式 $p \Rightarrow r$ は $Rest_1$ の要素でない(A)、または要素である(B)のいずれかである。

(A) $p \Rightarrow r$ が $Rest_1$ の要素でないとする。 $p \Rightarrow r$ は $Rule_1$ の要素であるかルールでないかいずれかである。 $Rule_1$ の要素であるとき、 $p \wedge q \Rightarrow r$ は極小ルールにならないので判定済になる。ルールでないとき、 $support$ 値の単調性より

```
mysql> select * from con2;
```

att1	val1	att2	val2	deci	deci_value
size	large	weight	light	price	high
size	large	weight	light	price	low

2 rows in set (0.00 sec)

```
mysql> select * from rule2;
```

att1	val1	att2	val2	deci	deci_value	support	accuracy
end_attrib	NULL	NULL	NULL	NULL	NULL	NULL	NULL

1 row in set (0.00 sec)

```
mysql> select * from rest2;
```

att1	val1	att2	val2	deci	deci_value	support	accuracy
size	large	weight	light	price	high	0.200	0.500
size	large	weight	light	price	low	0.200	0.500

2 rows in set (0.00 sec)

図4 2回目 while ループで得られるテーブル, $con2$, $rule2$, $rest2$. 実行例では $con2$ が $SubIMP_2$ を表す。 $con2$ の2つの含意式は $accuracy$ 値をみたさないために、引き続き $rest2$ に登録されている

$support(p \wedge q \Rightarrow r) < \alpha$ となり、やはり判定済になる。従って、 $p \wedge q \Rightarrow r$ を $SubIMP_2$ に追加する必要はない。

(B) $p \Rightarrow r$ が $Rest_1$ の要素であるとする。このとき、含意式 $q \Rightarrow r \in IMP_1$ は $Rest_1$ の要素でない(B1)、または要素である(B2)のいずれかである。

(B1) $q \Rightarrow r$ が $Rest_1$ の要素でないとき、(A)と同様に $p \wedge q \Rightarrow r$ は判定済になる。従って、 $p \wedge q \Rightarrow r$ を $SubIMP_2$ に追加する必要はない。

(B2) $q \Rightarrow r$ が $Rest_1$ の要素であるとき、これは同じ結論部を持つ含意式なので、 $support$ 値の要件をみたす場合に $p \wedge q \Rightarrow r$ を $SubIMP_2$ に追加する。

上記、(A) から (B) により、 $p \Rightarrow r \in Rest_1$, $q \Rightarrow r \in Rest_1$ から含意式 $\tau: p \wedge q \Rightarrow r$ を作り、この τ が $support$ 値の要件をみたす場合に $SubIMP_2$ に追加する。 $SubIMP_2$ は $support$ 値をみたし未判定であるすべての含意式の集合になる。

命題2は $SubIMP_2$ を $Rest_1$ から作る際の性質を示している。ここで、含意式 $p \Rightarrow r$ と $q \Rightarrow s$ をそれぞれ頻出アイテム集合 $\{p, r\}$ と $\{q, s\}$ に対応させ、例2のアプリアリ法を通して命題2を考えてみる。頻出アイテム集合の候補は $\{p, r, q\}$, $\{p, r, s\}$, $\{q, s, p\}$, $\{q, s, r\}$ となるが、DIS-アプリアリでは結論部は1個なので、 $r = s$ に限り要素数3個の集合を生成でき、候補 $\{p, q, r\}$ ($p \wedge q \Rightarrow r$) を得る。その後、 $support$ 値の判定を行い $SubIMP_2$ を決めることができる。DIS-アプリアリは頻出アイテム集合を確認できた時点で対応する含意式の $accuracy$ 値を計算し、ルールになるか、ルールにならないか、未判定であるか区別する。未判定となった頻出アイテム集合に対して、再度、要素数が1増えた頻出アイテム集合を作る。

命題2を利用し、2回目の while ループでは $SubIMP_2$ に対して初回と同様な手続きを行う(図4)。しかし、 $SubIMP_3$ については命題2のような表現は難しい。各 $p \wedge q \Rightarrow s \in Rest_2$ の条件部にデスクリプタ r (図3の表 $rest1$ に登録されている含意式 $r \Rightarrow s$ で p や q の属性と異なる)を追加し、 $p \wedge q \wedge r \Rightarrow s \in IMP_3$ を作る。本含意式が $support$ 値の要件をみたし、得られている

ルールに冗長でない場合に $SubIMP_3$ に追加する. $SubIMP_n$ ($n \geq 4$) 以降も同様の手続きを行う. これらの手順により, 初めてルールと判定された含意式 τ (自動的に極小ルールになる) は極小ルールとして登録され, τ の冗長な含意式は $SubIMP_i$ に追加されない. 一方, 未判定な含意式は逐次判定され, 実質 $\bigcup_{i=1}^n IMP_i$ をすべて調べたことになる. つまり, 極小ルールの見落としは無く, 次の命題を得る.

命題 3 $DIS \psi$ において, α と β によって決まる極小ルールの集合に対して図 1 の $Rule = \bigcup_{i=1} Rule_i$ はすべての極小ルールの集合と一致する. つまり, DIS -アプリオリは健全かつ完全である.

注意 3 DIS -アプリオリは基本的に Agrawal のアプリオリ法を DIS に対応させているため, 計算の複雑さはアプリオリ法と同等になると考えられる. また, α 値は実行に強く影響する. 表 1 の実行では, α 値が小さいために冗長な含意式の足切りができず, 図 4 の $Rest_2$ のように候補がそのまま残る. $\alpha = 0$ にすると Pawlak や Skowron が定義するルール生成を行い (実際の処理手順は異なる), NP 困難な問題であることからサイズが大きい表データにおける $\alpha = 0$ の設定は不都合になる. DIS -アプリオリはある程度大きい α 値を指定する場合 (出現頻度が大きい含意式をルールにする場合), 効果的にルールを生成する.

注意 4 著者らは DIS -アプリオリ単独で枠組みを提案せず, DIS -アプリオリを NIS -アプリオリ構築のための前段階として提案する. DIS -アプリオリの単独の枠組みは, 表データのアプリオリ法として既に提案されている CBA アルゴリズム (Classification based on Associations) [36] や $APRIORI-C$ [37] と同等の枠組みになると考えられる.

4. 非決定情報表におけるラフ集合とルール生成

本節では DIS に情報の不完全性を追加する目的で提案された NIS の枠組みと NIS におけるルール生成に言及する.

4.1 非決定情報表の研究背景

情報の不完全性を扱う枠組みとして, Lipski の不完全情報データベース [3] (表 4) や Orłowska と Pawlak による非決定情報システム (Nondeterministic Information System, NIS) [4] (表 5) が提案されている.

Lipski の研究も Orłowska の研究も情報の不完全性を場合分けの問題にし, 表データに対する問い合わせ機能の強化を目指していたと考えられる. 我々は, 問い合わせではなく NIS からのルール生成を進めている. 非決定情報システム $NIS \Phi$ を次の 4 つ組で扱う $\Phi = (OB, AT, VAL_A (A \in AT), g)$ [4],

$g : OB \times AT \rightarrow P(\bigcup_{A \in AT} VAL_A)$ ($P(\bigcup_{A \in AT} VAL_A)$ はべき集合). 写像 g は対象 x と属性 A に集合 $g(x, A) \subseteq VAL_A$ を割り当て, $g(x, A)$ のいずれかの値は真であるが特定できないと解釈する. 特に, $g(x, A) = VAL_A$ である場合には一切の情報を得ていないことと見なす. この解釈により, Codd の空値

表 4 Lipski の不完全情報データベースの例 [3]. 対象 x_2 の年齢は 52, 53, 54, 55 または 56 であるが情報の不完全性により特定できない

OB	age	$dept\#$	$hireyear$	$salary$
x_1	[60, 70]	{1, ..., 5}	{70, ..., 75}	{10000}
x_2	[52, 56]	{2}	{72, ..., 76}	(0, 20000]
x_3	{30}	{3}	{70, 71}	(0, ∞)
x_4	(0, ∞)	{2, 3}	{70, ..., 74}	{22000}
x_5	{32}	{4}	{75}	(0, ∞)

表 5 Orłowska の非決定情報表の例 [4]. D_i は病名, a_1 と a_2 は症状の種類, v_i や u_j は症状の状況

OB	a_1	a_2
D_1	{ v_1, v_3 }	{ u_1, u_2, u_3 }
D_2	{ v_2, v_5 }	{ u_1 }
D_3	{ v_1, v_3, v_4 }	{ u_1, u_2 }
D_4	{ v_1 }	{ u_1, u_2 }
D_5	{ v_1, v_3 }	{ u_1 }
D_6	{ v_5 }	{ u_1 }

(null value) や Grzymała-Busse の欠損値を NIS で扱うことができる.

$NIS \Phi$ において, 各 $g(x, A)$ から 1 つの属性値 $v_{x,A}$ を選択し $g(x, A)$ を $v_{x,A}$ に置きかえると 1 つの DIS を得る. 著者らはこの DIS を $NIS \Phi$ から派生する DIS とよぶ. 一般に複数の派生する DIS が存在するので, その集合を $DD(\Phi)$ とかく. 例えば, 表 5 では 288 ($= 6 \times 2 \times 6 \times 2 \times 2 \times 1$) 個の派生する DIS が存在し, NIS の定義から $DD(\Phi)$ の中には真の情報を有する $DIS \psi^{actual}$ が存在すると考える.

情報が不確実な場合, 一般に, 確実に成立すること (例えば選挙における開票時には不確実性が残るため, 当選ではなく当選確実と報道される), 可能性があること (例えば台風の進路予想では予報円や強風域により影響の可能性を示す), それ以外に分類することは自然である. Yao らもこの方針を three-way decisions [12] として取り上げている. この自然な方針により著者らは NIS におけるルールを次のように与える.

定義 5 [23, 26] (NIS におけるルール) $NIS \Phi$ において, 決定属性 Dec , $DD(\Phi)$, 指標値 α と β ($0 < \alpha, \beta \leq 1.0$) を定める.

(1) 含意式 τ が各 $\psi \in DD(\Phi)$ においてルールであるとき (各 ψ で $support(\tau) \geq \alpha$ と $accuracy(\tau) \geq \beta$ のとき), τ を Φ の確実ルールとよぶ.

(2) 含意式 τ が少なくとも 1 つの $\psi \in DD(\Phi)$ においてルールであるとき (少なくとも 1 つの ψ で $support(\tau) \geq \alpha$ と $accuracy(\tau) \geq \beta$ のとき), τ を Φ の可能性ルールとよぶ.

定義 5 は様相論理の立場からも極自然な定義であるが, $DD(\Phi)$ の要素数に依存する処理が必要であり研究的には取り上げられずにそのままになっていたと考えられる. 以下の節において, 定義 5 のルール生成に対する一解法を概説する.

4.2 様相性を考慮したルールとルール生成の基本定理

DIS においてデスクリプタ $[A, v]$ の同値類を用いるように、NIS においては下記の集合 inf と sup を考える。

定義 6 [23, 26] NIS Φ における各デスクリプタ $[A, v]$ について、以下の集合 inf と sup を定義する。

- (1) $inf([A, v]) = \{x \in OB \mid g(x, A) = \{v\}\}$,
- (2) $inf(\wedge_i [A_i, v_i]) = \cap_i inf([A_i, v_i])$,
- (3) $sup([A, v]) = \{x \in OB \mid v \in g(x, A)\}$,
- (4) $sup(\wedge_i [A_i, v_i]) = \cap_i sup([A_i, v_i])$.

一般に、 $inf([A, v]) \subseteq eq([A, v]) \subseteq sup([A, v])$ ($eq([A, v])$ は任意の $\psi \in DD(\Phi)$ における同値類) が成立し、 $inf([A, v])$ は情報が確定した対象の集合、 $sup([A, v]) \setminus inf([A, v])$ は情報の変動が可能な対象の集合を意味する。

定義 7 [23, 26] NIS Φ における含意式 τ について以下の指標を定義する。

- (1) $minsupp(\tau) = \min_{\psi \in DD(\Phi)} \{\psi \text{ での } support(\tau)\}$,
- (2) $minacc(\tau) = \min_{\psi \in DD(\Phi)} \{\psi \text{ での } accuracy(\tau)\}$,
- (3) $maxsupp(\tau) = \max_{\psi \in DD(\Phi)} \{\psi \text{ での } support(\tau)\}$,
- (4) $maxacc(\tau) = \max_{\psi \in DD(\Phi)} \{\psi \text{ での } accuracy(\tau)\}$.

定義 7 は $DD(\Phi)$ に依存し、Mammographic Mass や Congressional Voting データでは $DD(\Phi)$ の要素数が 10^{100} を超える。指標値の計算は困難と考えられるが、定義 6 の性質を利用すれば計算できる。ここでは、結果のみ列挙する [23, 26]。

命題 4 [23, 26] NIS Φ における含意式 $\tau : \wedge_{A \in CON} [A, val_A] \Rightarrow [Dec, v]$ について下記が成立する。

(1) $DD(\Phi)$ には少なくとも 1 つの $\psi_{min} \in DD(\Phi)$ が存在し、 ψ_{min} における $support(\tau)$ は $minsupp(\tau)$ と一致し、 $accuracy(\tau)$ は $minacc(\tau)$ と一致する。その時の指標値は、 inf , sup と $OUTACC$ を用いて以下のように表現される。

$$\begin{aligned} minsupp(\tau) &= \frac{|inf(\wedge_{A \in CON} [A, val_A]) \cap inf([Dec, v])|}{|OB|}, \\ minacc(\tau) &= \frac{|inf(\wedge_{A \in CON} [A, val_A]) \cap inf([Dec, v])|}{|inf(\wedge_{A \in CON} [A, val_A])| + |OUTACC|}, \\ OUTACC &= \{sup(\wedge_{A \in CON} [A, val_A]) \setminus inf(\wedge_{A \in CON} [A, val_A])\} \setminus inf([Dec, v]). \end{aligned}$$

(2) $DD(\Phi)$ には少なくとも 1 つの $\psi_{max} \in DD(\Phi)$ が存在し、 ψ_{max} における $support(\tau)$ は $maxsupp(\tau)$ と一致し、 $accuracy(\tau)$ は $maxacc(\tau)$ と一致する。その時の指標値は、 inf , sup と $INACC$ を用いて以下のように表現される。

$$\begin{aligned} maxsupp(\tau) &= \frac{|sup(\wedge_{A \in CON} [A, val_A]) \cap sup([Dec, v])|}{|OB|}, \\ maxacc(\tau) &= \frac{|sup(\wedge_{A \in CON} [A, val_A]) \cap sup([Dec, v])|}{|inf(\wedge_{A \in CON} [A, val_A])| + |INACC|}, \\ INACC &= \{sup(\wedge_{A \in CON} [A, val_A]) \setminus inf(\wedge_{A \in CON} [A, val_A])\} \cap sup([Dec, v]). \end{aligned}$$

(3) inf と sup を用いて指標値を計算することができ、この計

図 5 点 $(support(\tau), accuracy(\tau))$ の分布

算は $DD(\Phi)$ の要素数には依存しない。

図 5 [26] は命題 4 の内容をまとめたものである。どの含意式 τ についても点 $(support(\tau), accuracy(\tau))$ は $(minsupp(\tau), minacc(\tau))$ と $(maxsupp(\tau), maxacc(\tau))$ を頂点とする長方形の区画に位置する。一般に、

$$\begin{aligned} minsupp(\tau) &\leq \text{各 } \psi \text{ における } support(\tau) \leq maxsupp(\tau), \\ minacc(\tau) &\leq \text{各 } \psi \text{ における } accuracy(\tau) \leq maxacc(\tau) \end{aligned}$$

であるために、図 5 の特徴的な ψ_{min} と ψ_{max} 2 点を調べることでルールか否かの判定が可能になる。

定理 1 (NIS のルール生成における基本定理) [23, 26] NIS Φ において α と β を定める。含意式 τ について下記が成立する。

- (1) τ が確実ルールであるための必要十分条件は、 τ が ψ_{min} でルールになることである。つまり、 $minsupp(\tau) \geq \alpha$ と $minacc(\tau) \geq \beta$ が成立することである。
- (2) τ が可能性ルールであるための必要十分条件は、 τ が ψ_{max} でルールになることである。つまり、 $maxsupp(\tau) \geq \alpha$ と $maxacc(\tau) \geq \beta$ が成立することである。

定理 1 と命題 4 より、含意式 τ を 1 つ定めればそれが NIS においてルールになるか否かの判定ができることになる。そして、定理 1 は第 3 節で示した DIS-アプリオリアルゴリズムとの親和性がある。DIS-アプリオリではルールの候補になる含意式を逐次列挙しながら指標値の判定を行い、 $support$ 値と $accuracy$ 値の要件をみたすすべての極小ルールを生成する。従って、 $support$ 値を $minsupp$ 値、 $accuracy$ 値を $minacc$ 値に置き換えればすべての極小な確実ルールが生成される (図 6)。同様に、 $support$ 値を $maxsupp$ 値、 $accuracy$ 値を $maxacc$ 値に置き換えればすべての極小な可能性ルールが生成される (図 7)。図 6 と図 7 のアルゴリズムを合わせて、NIS-アプリオリ (アルゴリズム) とよぶ。

定理 2 [23, 26] NIS Φ において α と β を定める。 α と β により定まる極小な確実ルールの集合 (図 6) について NIS-アプリオリは健全かつ完全であり、極小な可能性ルールの集合 (図 7) について NIS-アプリオリは健全かつ完全である。

定理 2 はルール生成アルゴリズムに 1 つの妥当性を与えてい

Require: NIS Φ , 決定属性 Dec , 閾値 α, β .

Ensure: 極小な確実ルールの集合 $CERTAIN_Rule = \cup_{i=1} Rule_i$.

```

1:  $i \leftarrow 1$ ;
2:  $SubIMP_i = \{\tau_{i,j} \in IMP_i \mid \underline{minsupp}(\tau_{i,j}) \geq \alpha\}$  を作る;
3: while ( $|SubIMP_i| \geq 1$ ) do
4:    $Rest_i \leftarrow \{\}; Rule_i \leftarrow \{\};$ 
5:   for all  $\tau_{i,j} \in SubIMP_i$  do
6:     if  $\underline{minacc}(\tau_{i,j}) \geq \beta$  then  $\tau_{i,j}$  を  $Rule_i$  に追加;
7:     else  $\tau_{i,j}$  を  $Rest_i$  に追加;
8:   end if
9: end for
10:   $i \leftarrow i + 1$ ;
11:   $Rest_{i-1}$  を用い  $SubIMP_i = \{\tau_{i,j} \in IMP_i \mid \underline{minsupp}(\tau_{i,j}) \geq \alpha\}$ 
    を作る;
12:  を作る;
13: end while
14: return  $CERTAIN\_Rule = \cup_{i=1} Rule_i$ 

```

図 6 NIS-アプリアリアルゴリズム (確実ルール生成). 下線部が DIS-アプリアリアルゴリズムから変更されている

Require: NIS Φ , 決定属性 Dec , 閾値 α, β .

Ensure: 極小な可能性ルールの集合 $POS_Rule = \cup_{i=1} Rule_i$.

```

1: NIS-アプリアリアルゴリズム (確実ルール生成) の  $\underline{minsupp}$  を
   $\underline{maxsupp}$ ,  $\underline{minacc}$  を  $\underline{maxacc}$  に置き換えたアルゴリズムにする;
2: return  $POS\_Rule = \cup_{i=1} Rule_i$ 

```

図 7 NIS-アプリアリアルゴリズム (可能性ルール生成)

と考える。次に、NIS-アプリアリの計算量に言及するが、詳細な解析は難しい。定義 7 より、指標値 $\underline{minsupp}$ と $\underline{minacc}$ や $\underline{maxsupp}$ と $\underline{maxacc}$ は $DD(\Phi)$ に依存するが、それらは命題 4 により $DD(\Phi)$ とは無関係に $\inf$ と $\sup$ の演算で計算できる。この計算は定義 1 の同値類 $eq()$ による $\underline{support}$ や $\underline{accuracy}$ の計算とほぼ同等と考えられる。NIS-アプリアリでは確実ルール生成と可能性ルール生成で 2 度 DIS-アプリアリを用いているため、DIS-アプリアリの 2 倍の計算量が必要になる。注意 3 で言及したように DIS-アプリアリは Agrawal のアプリアリ法と同等の計算量と考えられるので、NIS-アプリアリの計算量は Agrawal のアプリアリ法の 2 倍程度の計算量になると考える。つまり、NIS-アプリアリは $DD(\Phi)$ (指数オーダーで増加する) に依存する計算を効率的とされる Agrawal のアプリアリ法の 2 倍程度の計算量で実行すると考えられ、十分に効率的なアルゴリズムであると考えられる。

次に、NIS-アプリアリによる Mammographic Mass データ (対象数 961, 属性数 6, $Dec = severity$, $\alpha = 0.2$, $\beta = 0.8$, $|DD(\Phi_{mammo})| \geq 10^{100}$) の実行例を示す (図 8, 図 9)。確実ルール ($c1_rule$, $c2_rule$) は 10^{100} 個を超える $\psi \in DD(\Phi_{mammo})$ のいずれにおいてもルールである含意式、可能性ルール ($p1_rule$, $p2_rule$) は 10^{100} 個を超える $\psi \in DD(\Phi_{mammo})$ の中の少なくとも 1 つの ψ でルールになる含意式であり、命題 4, 定理 1, 定理 2 が無ければ計算は難しい。その他にも実行例をウェブ上に公開している [29]。

```

mysql> show tables;
+-----+
| Tables_in_nis(mammo2) |
+-----+
| nrdf                    |
| table 1                 |
+-----+
2 rows in set (0.00 sec)

mysql> call step1('severity',961,0.2,0.8);
Query OK, 0 rows affected (13.20 sec)

mysql> call step2('severity',961,0.2,0.8);
Query OK, 0 rows affected (14.14 sec)

mysql> call step3('severity',961,0.2,0.8);
Query OK, 0 rows affected (7.00 sec)

mysql> show tables;
+-----+
| Tables_in_nis(mammo2) |
+-----+
| c1_rule                 |
| c2_rule                 |
| c3_rule                 |
| condi                  |
| nrdf                    |
| p1_rule                 |
| p2_rule                 |
| p3_rule                 |
| table 1                 |
+-----+
9 rows in set (0.01 sec)

```

図 8 ルール ($\underline{support}(\tau) \geq 0.2, \underline{accuracy}(\tau) \geq 0.8$) の生成. $|DD(\Phi_{mammo})| \geq 10^{100}$ の場合分けに相当する計算を命題 4, 定理 1 の手法で行っている

```

mysql> select * from c1_rule;
+-----+-----+-----+-----+-----+-----+
| att1 | val1 | deci | deci_value | minsupp | minacc |
+-----+-----+-----+-----+-----+-----+
| assess | 5 | severity | 1 | 0.317 | 0.882 |
| margin | 1 | severity | 0 | 0.329 | 0.859 |
| end_attrib | NULL | NULL | NULL | NULL | NULL |
+-----+-----+-----+-----+-----+-----+
3 rows in set (0.00 sec)

mysql> select * from p1_rule;
+-----+-----+-----+-----+-----+-----+
| att1 | val1 | deci | deci_value | maxsupp | maxacc |
+-----+-----+-----+-----+-----+-----+
| age | m | severity | 0 | 0.228 | 0.820 |
| age | o | severity | 1 | 0.266 | 0.892 |
| assess | 5 | severity | 1 | 0.318 | 0.884 |
| margin | 1 | severity | 0 | 0.367 | 0.886 |
| shape | 1 | severity | 0 | 0.213 | 0.844 |
| shape | 2 | severity | 0 | 0.203 | 0.848 |
| end_attrib | NULL | NULL | NULL | NULL | NULL |
+-----+-----+-----+-----+-----+-----+
7 rows in set (0.00 sec)

mysql> select * from c2_rule;
+-----+-----+-----+-----+-----+-----+
| att1 | val1 | att2 | val2 | deci | deci_value | minsupp | minacc |
+-----+-----+-----+-----+-----+-----+
| end_attrib | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
+-----+-----+-----+-----+-----+-----+
1 row in set (0.00 sec)

mysql> select * from p2_rule;
+-----+-----+-----+-----+-----+-----+
| att1 | val1 | att2 | val2 | deci | deci_value | maxsupp | maxacc |
+-----+-----+-----+-----+-----+-----+
| age | s | assess | 4 | severity | 0 | 0.250 | 0.870 |
| age | s | shape | 4 | severity | 1 | 0.212 | 0.864 |
| end_attrib | NULL | NULL | NULL | NULL | NULL | NULL | NULL |
+-----+-----+-----+-----+-----+-----+
3 rows in set (0.00 sec)

```

図 9 得られた極小な確実ルールと極小な可能性ルール. NIS-アプリアリの完全性から、これ以外にルールはないと保証される

5. まとめ

本論文では、過去のルール生成に関連する重要な研究に言及し、著者らが進めている NIS からのルール生成について述べた。可能世界意味論に基づくルールの定義は自然であるが、派生する DIS の個数が 10^{100} を超える場合もあり、今までにこの

体系の研究は成されていないと思われる。図9のようなルール生成は著者らの研究により初めて可能になったと考える。著者らの枠組みの特徴を再度、列挙する。

(1) 著者らはラフ集合の同値類の手法により、集合

$$\{inf([A, val_A]) \mid A \in AT, val_A \in VAL_A\},$$

$$\{sup([A, val_A]) \mid A \in AT, val_A \in VAL_A\}$$

を定義し、派生する DIS に基づく指標値の計算 (通常、指数オーダー) を $\{inf([A, val_A])\}$ と $\{sup([A, val_A])\}$ の処理 (命題4) で計算できることを示した。

(2) (1) の性質を Agrawal のアプリオリ法に組み込み、DIS-アプリオリと NIS-アプリオリを提案している。DIS-アプリオリは Agrawal のアプリオリ法と同等の計算量、工夫が無ければ指数オーダーになる NIS-アプリオリは DIS-アプリオリの2倍程度の計算量になると考える。情報の不完全性も考慮でき、計算量も十分に押さえられる。これにより、実際のルール生成における NIS-アプリオリの利用が可能になると考えられる。

(3) DIS-アプリオリ、NIS-アプリオリとも極小ルールの集合に対して健全かつ完全である。閾値 α と β で決まる極小ルールの集合に対して、過不足なくルールを生成する。実行時間は閾値 α によって調整可能である。

(4) プログラムの改善は必要な状況であるが、提案した NIS-アプリオリは SQL により実装されており、実際にルール生成が可能なる状況にある。

おわりに、ルール生成以外のラフ集合非決定情報解析 (RNIA) の研究の可能性を示す (図10)。著者らは DIS と NIS におけるルール生成の研究を進めて来たが、図10では NIS から DIS を推定したり、情報を意図的に曖昧化する枠組みも考慮できると考える。図10の (A) と (B) について簡単に述べる。

(A) ルール生成による機械学習：NIS-アプリオリはルール生成時の副産物として ψ_{min} や ψ_{max} を得ている。得られた含意式が十分に信用できる場合、NIS Φ で属性値が確定しない値を ψ_{max} の値で補うことで、NIS の値を部分的に特定する。これを繰り返し、NIS の確定していない値の推定を行うことができると考える。統計における最尤推定は得られたサンプルが最も生じやすくなるようにパラメータの推定を行う。これを真似て、NIS ではルールとして得られた含意式がさらに生じやすくなるように NIS の値の特定を行う。簡単のために、この枠組みをルール生成による機械学習 [26, 38] とよび応用の可能性を検討している。

(B) プライバシー保護への応用：NIS は情報の不完全性を扱う体系として提案されている。表4のような表現を認めれば、各 $x_i \in OB$ の個人情報情報は曖昧化される。例えば、三者択一 (P, Q, R のいずれかを選択) のアンケートにおいて回答者に P または Q 等の回答を認めれば回答者自身でプライバシーを保護できる。このようなアンケートで得られる情報表は DIS ではなく NIS になる。NIS-アプリオリはこのようなアンケート結果からもルールを生成できる [39]。

また、実データへの具体的応用として岩質データ解析 (Lithologic data analysis) の研究 [40] を始めている。データ解析により岩質データの特徴を把握できたり、地下資源を効率よく

DIS, NIS, Data mining in various types of data, Estimation of the actual value, Machine learning by rule generation, Information dilution, and Privacy-preserving issue

図10 RNIA と関連する研究課題 [38]

発見するための基礎知識やルールを得ることができれば望ましい。元データはあいまいな情報になることが想定されるため、RNIA が有効に活用できないかと考える。並行して、ビッグデータへの応用も進めている [27]。

表データにおける情報の不完全性と表データにおけるルール生成の観点から重要と思われる研究、著者らの研究を述べた。この分野に興味を持たれている皆様の参考になれば幸いと考える。

謝辞

本研究は JSPS 科研費, JP22500204, JP26330277, JP20K11954 の助成を受けたものです。

参考文献

- [1] W. Marek and Z. Pawlak: "Information storage and retrieval systems: Mathematical foundations," *Theoretical Computer Science*, Vol.1, No.4, pp. 331-354, 1976.
- [2] E. F. Codd: "A relational model of data for large shared data banks," *Communication of the ACM*, Vol.13, No.6, pp. 377-387, 1970.
- [3] W. Lipski: "On semantic issues connected with incomplete information databases," *ACM Trans. on Database Systems*, Vol.4, No.3, pp. 262-296, 1979.
- [4] E. Orłowska and Z. Pawlak: "Representation of nondeterministic information," *Theoretical Computer Science*, Vol.29, No.1-2, pp. 27-39, 1984.
- [5] Z. Pawlak: "Rough sets," *Int. J. of Computer and Information Sciences*, Vol.11, No.5, pp. 341-356, 1982.
- [6] Z. Pawlak: *Rough sets: Theoretical aspects of reasoning about data*, Kluwer Academic Publishers, 1991.
- [7] S. Greco, B. Matarazzo, and R. Słowiński: "Granular computing and data mining for ordered data: The dominance-based rough set approach," in *Encyclopedia of Complexity and Systems Science*, R. A. Meyers ed., Springer, pp. 4283-4305, 2009.
- [8] J. Komorowski, Z. Pawlak, L. Polkowski, and A. Skowron: "Rough sets: a tutorial," in *Rough Fuzzy Hybridization: A New Method for Decision Making*, S. K. Pal and A. Skowron eds., Springer, pp. 3-98, 1999.
- [9] Y. H. Qian, J. Y. Liang, Y. Yao, and C. Y. Dang: "MGRS: A multi-granulation rough set," *Information Sciences*, Vol.180, No.6, pp. 949-970, 2010.
- [10] A. Skowron and C. Rauszer: "The discernibility matrices and functions in information systems," in *Intelligent Decision Sup-*

port: *Handbook of Advances and Applications of the Rough Set Theory*, R. Słowiński ed., Kluwer Academic Publishers, pp. 331-362, 1992.

- [11] S. Tsumoto: “Automated extraction of hierarchical decision rules from clinical databases using rough set model,” *Expert Systems with Applications*, Vol.24, No.2, pp. 189-197, 2003.
- [12] Y. Y. Yao: “Three-way decisions with probabilistic rough sets,” *Information Sciences*, Vol.180, Vol.3, pp. 314-353, 2010.
- [13] W. Ziarko: “Variable precision rough set model,” *J. of Computer and System Sciences*, Vol.46, No.1, pp. 39-59, 1993.
- [14] 中村昭: “ラフ集合—その理論と応用,” 日本ファジイ学会誌, Vol.8, No.4, pp. 594-603, 1996.
- [15] 津本周作: “ラフ集合論の現状と課題,” 日本ファジイ学会誌, Vol.13, No.6, pp. 552-561, 2001.
- [16] 乾口雅弘: “ラフ集合の一般化—類似関係, ファジイ関係, 順序関係の下でのラフ集合—,” 日本ファジイ学会誌, Vol.13, No.6, pp. 562-570, 2001.
- [17] 村井哲也: “ラフ集合と様相論理,” 日本ファジイ学会誌, Vol.13, No.6, pp. 571-580, 2001.
- [18] 工藤康生: “ラフ集合のこれまでとこれから,” 知能と情報, Vol.30, No.4, pp. 209-211, 2018.
- [19] 楠木祥文: “決定ルール抽出の基礎: 被覆アルゴリズム,” 知能と情報, Vol.30, No.4, pp. 205-208, 2018.
- [20] J. W. Grzymała-Busse and P. Werbrück: “On the best search method in the LEM1 and LEM2 algorithms,” in *Incomplete Information: Rough Set Analysis, Studies in Fuzziness and Soft Computing*, E. Orłowska ed., Springer, pp. 75-91, 1998.
- [21] J. W. Grzymała-Busse and W. Rząsa: “A local version of the MLEM2 algorithm for rule induction,” *Fundamenta Informaticae*, Vol.100, No.1-4, pp. 99-116, 2010.
- [22] M. Kryszkiewicz: “Rules in incomplete information systems,” *Information Sciences*, Vol.113, No.3-4, pp. 271-292, 1999.
- [23] H. Sakai, M. Wu, and M. Nakata: “Apriori-based rule generation in incomplete information databases and non-deterministic information systems,” *Fundamenta Informaticae*, Vol.130, No.3, pp. 343-376, 2014.
- [24] A. Frank and A. Asuncion: UCI machine learning repository, School of Information and Computer Science, University of California, Irvine, <http://mllearn.ics.uci.edu/MLRepository.html> [accessed Jul. 10, 2019]
- [25] H. Sakai and A. Okuma: “Basic algorithms and tools for rough non-deterministic information analysis,” *Trans. on Rough Sets*, Vol.1, pp. 209-231, 2004.
- [26] H. Sakai, M. Nakata, and J. Watada: “NIS-Apriori-based rule generation with three-way decisions and its application system in SQL,” *Information Sciences*, Vol.507, pp. 755-771, 2020.
- [27] Z. Jian, H. Sakai, J. Watada, A. Roy, and M. Hassan: “An Apriori-based data analysis on suspicious network event recognition,” *Proc. IEEE Int. Conf. on Big Data*, pp. 5888-5896, 2019.
- [28] H. Sakai: “On rough set-based rule generation and Apriori-based rule generation: A comparison of them and a combination for effective rule generators,” *The Future of Rough Sets (pdf file)*, *Int. Rough Set Society Web Page*, 2019, <https://www.roughsets.org/roughsets/going-forward/> [accessed Aug. 5, 2019]
- [29] H. Sakai: Execution logs by RNIA software tools, <http://www.mns.kyutech.ac.jp/~sakai/RNIA> [accessed Jul. 10, 2019]
- [30] R. Agrawal and R. Srikant: “Fast algorithms for mining association rules in large databases,” *Proc. of the 20th Int. Conf. on Very Large Data Bases (VLDB’94)*, pp. 487-499, 1994.
- [31] M. Nakata, H. Sakai, and K. Hara: “Rough sets based on possible indiscernibility relations in incomplete information tables with continuous values,” *Proc. Int. Conf. on Rough Sets (IJCRS 2019)*, pp. 155-165, 2019.
- [32] N. Shan and W. Ziarko: “Data-based acquisition and incremental modification of classification rules,” *Computational Intelligence*, Vol.11, No.2, pp. 357-370, 1995.

- [33] H. S. Nguyen and S. H. Nguyen: “Rough sets and association rule generation,” *Fundamenta Informaticae*, Vol.40, No.4, pp. 383-405, 1999.
- [34] J. R. Quinlan: “Introduction of Decision Trees,” *Machine Learning*, Vol.1, pp. 81-106, 1986.
- [35] J. W. Lloyd: *Foundations of Logic Programming*, Springer, p. 124, 1984.
- [36] B. Liu, W. Hsu, and Y. Ma: “Integrating classification and association rule mining,” *Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD’98)*, pp. 80-86, 1998.
- [37] V. Jovanoski and N. Lavrač: “Classification rule learning with APRIORI-C,” *Proc. of the 10th Portuguese Conf. on Artificial Intelligence (EPIA 2001)*, pp. 44-51, 2001.
- [38] H. Sakai, M. Nakata, and Y. Y. Yao: “Pawlak’s many valued information system, non-deterministic information system, and a proposal of new topics on information incompleteness toward the actual application,” *Studies in Computational Intelligence*, Vol.708, pp. 187-204, 2017.
- [39] H. Sakai, C. Liu, M. Nakata, and S. Tsumoto: “A proposal of a privacy-preserving questionnaire by non-deterministic information and its analysis,” *Proc. IEEE Int. Conf. on Big Data*, pp. 1956-1965, 2016.
- [40] T. M. Hossain, J. Watada, Z. Jian, H. Sakai, S. Rahman, and I. A. Aziz: “Missing well log data handling in complex lithology prediction: An NIS apriori algorithm approach,” *Int. J. of Innovative Computing, Information and Control*, Vol.16, No.3, pp. 1077-1091, 2020.

(2019 年 8 月 28 日 受付)

(2019 年 10 月 15 日 採録)

[問い合わせ先]

〒804-8550 福岡県北九州市戸畑区仙水町 1 番 1 号

九州工業大学 工学研究院

酒井 浩

TEL: 093-884-3258

FAX: 093-884-3258

E-mail: sakai@mns.kyutech.ac.jp

— 著 者 紹 介 —

酒井 浩 [正会員]

1988 年九州大学大学院総合理工学研究科博士課程修了。(理学博士)。現在, 九州工業大学工学研究院教授。国際ラフ集合学会(シニアメンバ), IEEE 学会, 日本知能情報ファジイ学会, 日本数学会の会員。ラフ集合, データマイニング, 様相論理の体系, 不完全情報, アプリオリアルゴリズム等の研究に従事。

中田 典規 [正会員]

1985 年東北大学大学院理学研究科博士課程修了。(理学博士)。現在, 城西国際大学教授。国際ラフ集合学会(シニアメンバ), ACM 学会, AAAI 学会, IEEE 学会, ASL 学会, 日本知能情報ファジイ学会の会員。ラフ集合, データマイニング, ファジィデータベース, 様相論理, 不完全情報等の研究に従事。

わただ じゅんぞう
和多田 淳三 [正会員]

1983 年大阪府立大学大学院博士課程修了.
(工学博士). 早稲田大学名誉教授, マレーシ
アペトロナス大学教授. FIM (Forum for Inter-
disciplinary Mathematics, India) 学会長, IEEE
学会, 日本知能情報ファジィ学会の会員 (フェ
ロー). ファジィ集合, ラフ集合, 経営工学等
の研究に従事.

**Rough Sets Non-Deterministic Information Analysis and a NIS-Apriori System
– A Rule Generation System Based on Possible World Semantics –**

by

Hiroshi SAKAI, Michinori NAKATA, and Junzo WATADA

Abstract:

In this paper, information incompleteness in tabular data and rule generation in tabular data are focused on, and the previously proposed important researches are surveyed from the viewpoint of rough sets and data mining. Then, “Rough sets Non-deterministic Information Analysis (RNIA)” and “a NIS-Apriori system” proposed by the authors are described. The framework of rule generation from tables with information incompleteness is also outlined. RNIA offers a new framework of rough sets based on possible world semantics, and it gives one solution algorithm for the rule generation problem considering information incompleteness. This solution algorithm is termed “NIS-Apriori” and implemented as a NIS-Apriori system in SQL language. We also mention the new possibility of rough sets non-deterministic information analysis that employs the NIS-Apriori algorithm as the core algorithm.

Keywords: rule generation, rough sets, non-deterministic information, possible world semantics, Apriori algorithm

Contact Address: **Hiroshi Sakai**

1-1 Sensui, Tobata, Kitakyushu, Fukuoka 804-8550, Japan

TEL: +81-93-884-3258

FAX: +81-93-884-3258

E-mail: sakai@mns.kyutech.ac.jp