2A3-6

ルール群視察ガイドシステム
Rule groups Inspection Guidance System

中野 優 Yu Nakano　　岡田 孝 Takashi Okada

関西学院大学 理工学部 情報科学科
Department of Informatics, School of Science and Technology, Kwansei Gakuin University

The characteristic rule induction usually produces a large number of rules, and it is difficult for a user to inspect all rules. This paper describes a method to give a priority index to rules based on their supporting instances, and it guides a user to inspect the most useful rule successively. The priority index is calculated dynamically at each step of rule set inspection using the covered instances by the employed rules, and the resulting rule group obtained gives a concise understanding to the data of the target class. A GUI system has been developed to guide the survey of rules, which shows a SOM picture of rules that is useful to grasp the relationships among rules.

1. はじめに

今日，様々な分野で大量のデータから有用な知識を発掘するデータマイニングが行われている．その代表的な手法として，相関ルール導出[Agrawal 93, 94]が挙げられる．相関ルールとそれをさらに発展させたカスケードモデル[Okada 00]は特徴的ルール導出と呼ばれ，現在データマイニングの主要な技法の1つとなっている．

相関ルール導出の主要な強みは網羅性であり，ユーザが設定した最小支持度と最小確信度の制約を満たす，すべての相関関係を導き出す．しかし，この強みは同時に，非常に膨大なルールを出力するという大きな欠点も持っており，その数は数千から数万件になる場合もある．そのため，ユーザがルール全体を視察することが現実的に困難になる．この問題の対処策として，最低支持度や最低確信度の値を上げることにより，出力されるルール数を減らすことができるが，その結果は既知の内容に限られることが多く，解析自体が無意味となる．

本研究では，特徴的ルール導出により出力されたルール数は，必然的に多くなるものだという立場で，ルール視察方法に工夫を加える．既に視察したルール(既得ルール)群から得られる情報をもとに，まだ視察していないルール(候補ルール)群の中から有用な情報を提供するルールを各視察段階で動的に求め，有用であるルールを優先的に表示するシステムの構築を目的とする．その指標として，ルールの支持事例に基づいた優先度の導入を試みた．またルール視察を視覚的に補助するため，対話的な GUI(Graphical User Interface)の開発も併せて行った．

2. 方法

2.1 問題定義

候補ルールが満たす事例数（m）に対して，既得ルール群が満たす事例と重なる事例（Overlap），重ならない新規事例（New）に分け，さらにそれらを学習に用いられた目的変数のカテゴリ別（T, F）に分ける．同様に，候補ルールが満たさない事例（$\overline{m}$）に対しても同じ処理を行う．このようにして得られた度数表を表 1 に示す．ここで，各セルは候補ルールを（満たす事例数）／（満たさない事例数）を表記している．

表 1.　候補ルールに関する事例度数表

既得ルール ＼ 目的変数のクラス	T	F	Sum
New	m_{NT}／$\overline{m}_{NT}$	m_{NF}／$\overline{m}_{NF}$	m_{N}／$\overline{m}_{N}$
Overlap	m_{OT}／$\overline{m}_{OT}$	m_{OF}／$\overline{m}_{OF}$	m_{O}／$\overline{m}_{O}$
Sum	m_{T}／$\overline{m}_{T}$	m_{F}／$\overline{m}_{F}$	m／$\overline{m}$

候補ルール群を，図1のように支持事例数に基づいて図示し，ルールの特徴を視覚的に理解することを試みる．図全体の面積は，ルール導出に用いられたデータセット中のすべての事例を表現し，各部分の面積は，事例数に比例して大きくなる．また図の横軸は，学習に用いられた目的変数のカテゴリの数に分割され，分割比は，データセットの目的変数のカテゴリ分布に従うとする．図の縦軸は，既得ルール群を支持する事例と支持しない事例に分割する．さらに候補ルールを支持する事例を，既得ルール群を支持する事例と重複する部分，および重複しない部分に分割する．得られた図を候補ルール図と呼ぶ．

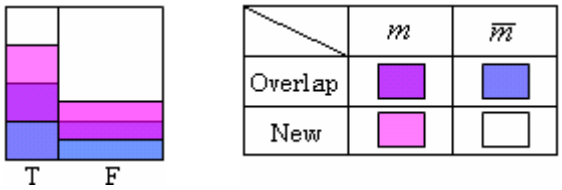

図 1.　候補ルール図

2.2 BSS

優先度を表現するために，Gini による平方和を用いる[Gini 12]．数値変数の平方和定義は(1)式のように変形できるが，ここでカテゴリ値の場合も事例 i, j 間での x_i-x_j の値を，x_i=x_j の時に 0，他は 1 とすれば，(2)式の平方和定義が得られる．ただし，n は全事例数を表し，$p(a)$はその属性が値 a をとる確率である．

$$SS = \frac{1}{2n}\sum_i \sum_j (x_i - x_j)^2 \qquad (1)$$

$$SS = \frac{n}{2}(1 - \sum_i p(a)^2) \qquad (2)$$

一群の事例をある属性の値で G 個の群に分割したとき，元の全平方和(TSS: Total sum of squares)は式(3)のようにそれぞれ

連絡先：岡田 孝，関西学院大学 理工学研究科 情報科学科，〒669-1337 兵庫県三田市学園 1-2，Tel：079-565-8149，okada-office@ksc.keansei.ac.jp

の群内平方和(WSS: Within group sum of squares)および群間平方和(BSS: Between groups sum of squares)に分割できる. BSS は式(4)で定義し, 添え字 U, L は分割前と後を指示する.

$$TSS = \sum_{g=1}^{G}(WSS_g + BSS_g) \tag{3}$$

$$BSS = \frac{n^L}{2}\sum_a (p_a^L(g) - p_a^U(g))^2 \tag{4}$$

2.3 優先度の設定

優先度を設定する際の重要な因子として新規度と強度を考慮し, それらを BSS により表現する.

強度は支持事例数が多く, 分布変化が大きいという尺度で定義する. すなわち, ルールの面白さを表現するパラメータとなる. U, L で指定されている事例群を, 元データ全体と候補ルールを支持する事例に対応させる. 表 1 の変数を用いた強度を式(5)に示す.

$$Strength = \frac{m}{2}\sum_a (\frac{m_a}{m} - \frac{m_a + \overline{m}_a}{m + \overline{m}})^2 \tag{5}$$

新規度はこれまでに採用されたルール群で支持されていない新規事例を, 新たな候補ルールがもつことに起因する. しかし, 新規事例をすべて新規度として評価すべきであろうか. もしも, ルール全体としての精度が高くとも, 新規事例群での精度が悪いルールを候補ルールとして採用すると, ルール群全体としての情報を劣化させる恐れがある. そこで, 新規度の定義としては, 新規事例群を式(4)の L とした場合の *BSS* 値を採用する. ただし, 注目する目的クラスとは異なった値を特徴づけるルールは, 候補ルールから除外する. 表1の変数を用いた新規度の定義を式(6)に示す. ただし X を目的クラスとする.

$$\begin{aligned} &if \quad \frac{m_{NX}}{m_N} - \frac{m_X + \overline{m}_X}{m + \overline{m}} < 0 \qquad Novelty = 0 \\ &else \quad Novelty = \frac{m_N}{2}\sum_a (\frac{m_{Na}}{m_N} - \frac{m_a + \overline{m}_a}{m + \overline{m}})^2 \end{aligned} \tag{6}$$

優先度は新規度と強度の積で表現し, 式(7)に示す. 新規度はルール視察が進むにつれて必然的に小さな値になる. すなわち, 既得ルール群と同じような事例を満たすルールの優先度は低くなり, 類似ルールを重複して視察することを防ぐことができる.

$$Priority = Strength \cdot Novelty \tag{7}$$

2.4 SOM

特徴的ルール導出は非常に多くのルールを出力するため, 得られた多くのルール間の関係や, その全体像を把握することが難しい. そこでルール群全体を見渡すため SOM(Self-Organizing Maps) 表示を導入する[Kohonen 96]. 入力ベクトルと重みの類似性マッチングはユークリッド距離を使用し, 近傍関数はトライアングル型の関数を用いた. 入力サンプル数は与えたルール数に等しく, またその属性値としてはデータセット中の各事例を支持するか否かとする. しかし, 全ルールに対して, 寄与していない事例は属性から削除する. ルール R_i, 事例番号 j に対して, 入力ベクトル x_{ij} の値は式(8)のように決定される.

$$x_{ij} = \begin{cases} 0 & if \quad j \notin R_i \\ 1 & if \quad j \in R_i \end{cases} \tag{8}$$

3. 結果と考察

3.1 実験データセットと視察過程

実験データセットとしては, MDL 社の MDDR データベース Ver.2001.1(23.01)に収載されているドーパミン受容体に対してアゴニスト活性を持つ化合物データセット(353 属性, 369 事例)を採用した. このデータセットからは, カスケードモデルにより D1 アゴニスト活性を目的変数とするルール群が得られている. このルール群 67 種の中で, 10 以上の支持化合物を持ち, さらに活性を有する化合物が 80%以上の 16 種のルールをあらかじめ選択して, 本研究における対象ルール群とした.

目的クラス[D1AGn:on]に対して, ルール優先度の高い順にルールを視察し, 候補ルールが無くなった時点で実験を終了する. また, SOM の初期学習パラメータを表 2 に示す.

表 2. SOM の初期学習パラメータ

parameter	value	parameter	value
学習回数	3000	初期学習率	0.8
出力ユニット数	36(6×6)	入力ベクトル数	16
初期学習範囲	4	入力ベクトル次元	100

3.2 ルール群の選択

DISCAS により出力されたルール数 16 個のうち, 本システムでは 3 つのルール Rule1, Rule1-UL8, Rule1-UL7 を順に視察することで実験終了条件が満たされた. 表 3 に各ルールの優先度, 強度, 新規度, 精度, 新規事例精度, 支持事例数, 各ルール視察段階における優先度の高い 5 つの候補ルールを示す. ただし, 精度, 新規事例精度, 支持事例数は目的クラスについてのみ記載する. また, ルールを選ぶ過程を, 各ルール視察段階の候補ルール図を用いて図 2 に示す.

表 3. 優先度により出力されたルール群の詳細

Inspection Order	Selected Rule	Parameter	Value	Nomination Rules
1	Rule1	優先度	1057.51	Rule1 Rule1-UL1 Rule1-UL6 Rule1-UL7 Rule1-UL8
		強度	32.51	
		新規度	32.51	
		精度	0.96	
		新規事例精度	0.96	
		支持事例数 [New, Overlap]	[50, 0]	
2	Rule1-UL8	優先度	73.77	Rule1-UL8 Rule1-UL7 Rule1-UL6 Rule1-UL10 Rule 14
		強度	17.87	
		新規度	4.12	
		精度	1	
		新規事例精度	1	
		支持事例数 [New, Overlap]	[6, 20]	
3	Rule1-UL7	優先度	8.45	Rule1-UL7 Rule1-UL6 Rule 14
		強度	19.49	
		新規度	0.43	
		精度	0.93	
		新規事例精度	0.5	
		支持事例数 [New, Overlap]	[2, 29]	

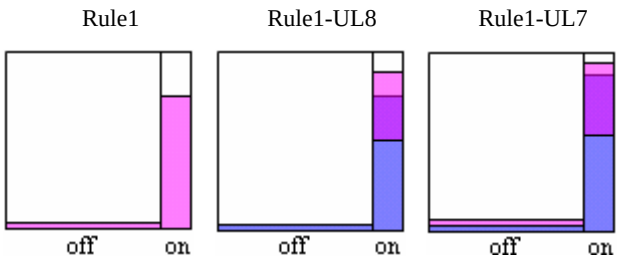

図 2.　各ルール視察段階の候補ルール図

表 3 の各視察段階で候補ルールの優先度は，既に選択したルールに依存して視察段階ごとに変化しており，動的に有用なルールを示していることが分かる．また，3 回目の視察では候補ルールは表 3 に示す 3 種に絞られている．また図 2 の既得ルール群と候補ルールの事例の重なり具合から，各ルール間の関係を視覚的に把握できる．Rule1 は，すべてのルールにより説明される目的クラス 58 事例の内，50 事例を説明している．Rule1-UL8，Rule1-UL7 が説明する事例のほとんどが Rule1 と重複しているが，Rule1 で説明されていない残りの事例を説明しており，また非常に強度が高いルールである．得られた既得ルール群は目的クラスのデータ全体を要約しているルール群であるといえる．

SOM の実行結果を図 3 に示す．青色の実線で囲まれたルールは，本システムにより得られたルール群，赤の点線で囲まれたルールは次節に述べる専門家ルール，また青色で塗られた部分は，本システムでの 2，3 回目の視察時の候補ルールを示している．候補ルールは下方に集中しており，その他の上部に位置するルールのほとんどが Rule1 と支持事例が重複しているルールであった．よって，SOM は支持事例とルールの関係をよく反映していると言える．

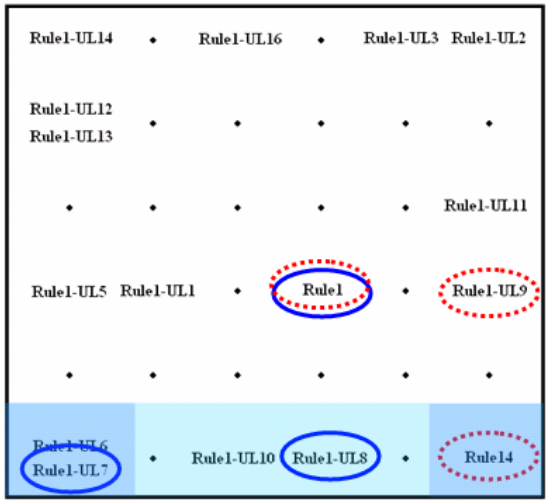

図 3　SOM の出力

3.3 専門家によるルール選択

実験に用いたルール群は，既に化学構造解析の専門家により，手作業で詳細に解析されている[Yamakawa 05]．その結果，興味深いルールとして選択されたのは，実験により得られたルール群とは異なり，Rule1，Rule1-UL9，Rule14 であった．これらのルールを専門家ルールと呼び，以下にこれらの相違について考察を加える．

Rule1-UL9 の支持事例は Rule1 の支持事例とすべて重複しており，Rule1-UL9 が表現する内容も Rule1 を視察することにより説明できる内容であった．ただし，Rule1-UL9 は Rule1 で表現される事例群中で，特定の骨格構造を有する化合物に特化した内容を表現しているため，専門家による解釈を加えやすいという理由から選択されている．一方 Rule14 も同じ理由で選択されたものであるが，この場合は新規事例も含むため，表 3 でも優先度が高い候補ルールとして表示されている．

専門家に実際に本システムを使用してもらい評価を依頼した．その結果，これまでのルール視察では，重複する事例を説明するルールの採択を避けるために注意を払う必要があったのに対し，このシステムを利用することにより，各ルールへの優先度設定がルール視察の方向性を決定できて有用であると評価された．

また多くのルール群を闇雲に視察するのではなく，SOM によりルール群の関係を把握した上で視察できるため，安心感を持って効率的な視察が行える点に高い評価を得ることができた．

実際，専門家ルールの Rule1-UL9 は，SOM 上で Rule1 と非常に近い位置に配置されている．したがって，Rule1 の表現で解釈が困難な場合に，その近傍のルール Rule1-UL9 の視察を試みてみればよいことになる．同様の選択は，Rule1-UL8 をシステムが優先度の高いルールとして示唆した場合にも行える．すなわち，このルールが専門家には解釈困難であっても，その近傍にありかつ優先度の高い Rule14 を選んで，容易に視察が行える．よって SOM はルール間の関係をよく反映しており，ルール視察に非常に有効であるといえる．

4. まとめ

本研究では，支持事例に基づいたルールの優先度設定により，同じ事例群を満たす類似ルールの重複した視察を回避でき，ルール選択の負担を大幅に軽減することができた．得られたルール群は，目的クラスのデータを特徴づける一種の要約となり，データセット全体を概観する際に有効であると考えられる．

また，ルール解釈に専門家の知識が必要な場合も，優先度と SOM を利用してユーザがルールを選択することにより，より柔軟かつ効率的なルール群の選択と視察による解析が可能となった．実際に，専門家による評価を依頼したところ，SOM によりルール群の関係を把握することができるため，ルール解析の負担が減少したことを確認した．

今後の課題として，優先度をより人間の感性に近づくように改良すること，またルール数が非常に多い場合に同様に有効であるか否かを確認することが挙げられる．

参考文献

[Agrawal 93] R. Agrawal, T. Imielinkski, and A. Swami: Mining Association Rules between Sets of Items in Large Databases, Proc. ACMSIGMOD pp.207-216 (1993).

[Agrawal 94] R. Agrawal: Fast Algorithms for Mining Association Rules, Proc. VLDB pp.487-499 (1994).

[Okada 00] Takashi Okada: Efficient Detection of Local Interactions in the Cascade Model, Knowledge Discovery and Data Mining , PAKDD-2000, LNAI 1805, pp.193-203.

[Gini 12] Gini C.W.: Variability and mutability, Studi Economico-Giuridici della R. Universita de Cagliari (1912).

[Kohonen 96] T. Kohonen: Self-Organizing Maps, Springer-Verlag Berlin Heidelberg New York (1996).

[Yamakawa 05] 山川眞透，岡田孝：”ドーパミン・アンタゴニストの特徴的構造について”，宝塚ワークショップ，アクティブマイニングによる化学構造からの知識発見, 26-32 (2005) .