1E4-3

ラフ集合と形式概念分析を用いた概念構造可視化システムの開発

Development of Conceptual Structure Visualization System using Rough Sets and Formal Concept Analysis

酒匂 一世[*1]　　原田 利宣[*1]
Issei Sakou　　Toshinobu Harada

[*1] 和歌山大学システム工学部デザイン情報学科
Department of Design and Information Sciences, Faculty of Systems Engineering, Wakayama University

Formal concept analysis (FCA) has a weak point that Hasse diagram is complicated when amount of attributes or samples large, and rough sets has a weak point that it is difficult to grasp structure of concept among several decision rules (DRs). Therefore, the purpose of this study was to combine two analytical methods in order to make up for the weak points of two methods mutually, and enable analysts to find relation of concepts that it is difficult to find by using two methods individually. Next, we developed a conceptual structure visualization system that executes two analytical methods simultaneously and projects DRs on Hasse diagram. Lastly, we verified usefulness of the system through comparing the results of analyses by using the system with them without the system by using data in papers on FCA or rough sets.

1. はじめに

データマイニングとは，収集した大量のデータを解析することで，項目間に見られる規則性や関連性などの利用可能な知識を抽出する技術の総称である．近年においては，インターネットへの接続が可能なコンピュータが一般家庭に普及したことによるネット上に蓄積されるデータ量の増大，およびコンピュータが高性能化したことによるデータ解析の容易化にともなって，データマイニングが利用される機会が増加し，そのニーズは大きく増大している．

そこで本研究では，既存のデータマイニングアルゴリズムをもとに，より使いやすく実用性の高い新たなアルゴリズムを提案し，また提案したアルゴリズムを用いてデータマイニングシステムを開発した．加えて，開発したシステムを使用して実際に分析を行い，既存のアルゴリズムによる分析結果と比較することでシステムの有用性を検証した．

2. 提案するアルゴリズムの概要

新たなデータマイニングアルゴリズムを提案するにあたり，本研究では既存のデータマイニング手法の中から，形式概念分析とラフ集合に着目した．形式概念分析とラフ集合は表 1 に示すような特徴を持っている．

表 1:形式概念分析とラフ集合の特徴

	形式概念分析	ラフ集合
分析対象のデータ	属性表（コンテクスト表）	決定表
得られる分析結果	ハッセ図	決定ルール（If-Then形式の式）
利点	すべての属性間の概念関係を視覚的に理解することが可能	ある決定クラスに対する属性間の因果関係を簡潔に表現することが可能
欠点	分析対象や属性の数が増加にともないハッセ図が加速度的に複雑化する	複数の決定ルール間での属性の関係性が把握しづらい

形式概念分析はハッセ図による属性間の包含関係を主にした概念構造の可視化が，またラフ集合は決定ルールにより，ある結論に対する属性間の因果関係の簡潔な表現が可能である．しかし一方で，形式概念分析は分析対象や属性数が多いと図が複雑化してしまい，またラフ集合は複数の決定ルール間の概念構造（包含関係）を把握しづらいという欠点がある．

本研究で提案するアルゴリズムは，形式概念分析であつかう属性表に決定属性を設定することで，同一の分析対象に対して形式概念分析とラフ集合解析を同時に実行する．その後，形式概念分析によって得られたハッセ図の上にラフ集合解析で得られた決定ルールを投影，描画することで，両方の分析結果を単一の図の上で視覚化できるようにした．

形式概念分析とラフ集合を組み合わせることで得られる利点として，上にあげたような各分析手法が持つ欠点を互いに補いあうことができる点，また形式概念分析あるいはラフ集合を単体で使用した場合には得ることが難しい知識の抽出が可能となる点が考えられる．

3. 概念構造可視化システムの開発

3.1 システムの概要

本研究で開発した概念構造可視化システムの動作フローを図 1 に示す．

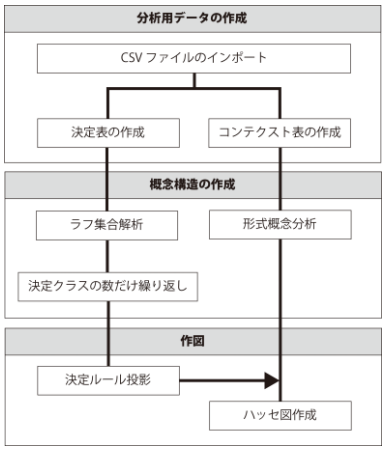

図 1:システムの動作フロー

連絡先:酒匂一世
和歌山大学システム工学部デザイン情報学科
〒640-8510 和歌山市栄谷 930 番地
E-mail:s175063@center.wakayama-u.ac.jp

本システムは，CSV 形式の決定表をインポートすることで分析を行う．インポートされた決定表はコンテクスト表に変換される．その後コンテクスト表に対して形式概念分析を，決定表に対してラフ集合解析を行い，ハッセ図と極小決定ルールを求めたのち，ウィンドウ上にハッセ図およびハッセ図に投影した極小決定ルールを表示する．

3.2 システムの機能

開発したシステムを用いて表 2 の決定表を分析した結果を図 2 に示す．決定表は書籍「Introduction To Lattices And Order」より抜粋した太陽系惑星の特徴データ[1]に決定属性「軌道傾斜角」を追加したものを用いた．表 2 における決定クラスは「軌道傾斜角」とした．図 2 を例として本システムの機能を説明する．

表 2：決定表（惑星の特徴データ）

	大きさ	太陽からの距離	衛星の有無	軌道傾斜角
水星	小	近い	無	大
金星	小	近い	無	小
地球	小	近い	有	小
火星	小	近い	有	小
木星	大	遠い	有	小
土星	大	遠い	有	小
天王星	中	遠い	有	小
海王星	中	遠い	有	小
冥王星	小	遠い	有	大

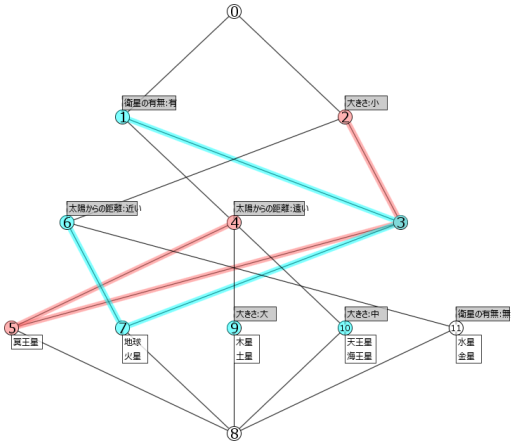

図 2：決定ルール投影ハッセ図（惑星の特徴データ）

(1) 基本機能

図 2 における灰色のウィンドウにはノードに対応する属性，白いウィンドウにはノードに該当するオブジェクトが表示されている．本システムで可能な操作の一部を下に示す．

1）ウィンドウごとの自由な表示・非表示切り替え
2）ドラッグによるノードや各ウィンドウの移動
3）マウスホイールによる図全体の拡大・縮小
4）マウスオーバーによるノードとそれに関連するウィンドウのハイライト表示
5）リンクの不透明度変更

ハッセ図上の各ノードは決定表における属性値とそれに属するサンプルをあらわしている．リンクはノードどうしの包含，被包含関係を示しており，図において上に位置しているノードが下に位置しているノードを包含している．
例えば，図 2 において属性値「大きさが小」に対応する 2 番ノードが属性値「太陽からの距離が近い」に対応する 6 番ノードの上位にあることから「太陽に近い惑星は必ず小さい」という属性間の関係を読みとることができる．

(2) ノードおよびリンクの着色

各ノードとリンクは決定クラスに応じて着色される．図 2 において赤く着色されている部分は決定クラス「軌道傾斜角：大」，青く着色されている部分は決定クラス「軌道傾斜角：小」の決定ルールである．例えば「軌道傾斜角：大」の決定ルールは If[大きさが小]and[太陽からの距離が遠い]Then[軌道傾斜角が大] となるので，図 2 においては，属性値「大きさ：小」と「太陽からの距離：遠い」に属するノードおよびそれらを結ぶリンクが赤く着色されている．

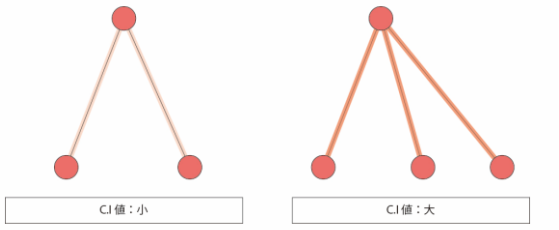

図 3：ハッセ図上における C.I 値の大小

また，着色される色の濃さは対応する決定ルールの C.I 値の大小によって変わり，図 3 に示すように C.I 値が大きい場合は濃い色で，小さい場合は薄い色で着色される．なお，異なる複数の決定ルールに含まれるものは，各決定ルールの色で重ね塗りされるため，ノードの色が混合される．

4. システムの評価

4.1 評価実験の概要

開発したシステムの有用性を検証するため，評価実験を行った．評価実験では，はじめに既存の論文や文献に掲載されている形式概念分析およびラフ集合用のデータに対し，開発したシステムを用いて分析を行う．次に，データが掲載されている論文や文献内における分析結果と，システムを用いた分析結果を比較し，双方の分析結果の違いを検討する．システムを用いた分析結果からのみ解釈が可能な情報があれば，本システムは今までのデータマイニングツールからは得られない情報を抽出することができ，有用であると言える．

次節より，評価実験を行った対象のデータのうち 2 つをあげ，それぞれの評価実験によって得られた結果を示す．

4.2 自動車の選好データを用いた比較

1 つめの検証用データとして，表 3 の自動車の選好データをあげる．これは，ラフ集合プログラムへの入力データ例として，書籍「ラフ集合の感性工学への応用」[2]において使用された決定表である．従来のラフ集合解析によって得られる決定ルールを表 4 に，本研究で開発したシステムから得られる決定ルール投影ハッセ図を図 3 に示す．決定属性は「選好」とした．

表 3：決定表（自動車の選好データ）

	カラー	造形	ドアタイプ	イメージ	選好
s1	色彩系	有機的	2 ドア	パーソナル	好き
s2	色彩系	曲線的	2 ドア	スポーティ	どちらでもない
s3	白黒系	曲線的	4 ドア	フォーマル	どちらでもない
s4	白黒系	有機的	4 ドア	パーソナル	好き
s5	白黒系	曲線的	4 ドア	パーソナル	どちらでもない
s6	色彩系	曲線的	2 ドア	スポーティ	好き

表 4:決定ルール(自動車の選好データ)

決定クラス	決定ルール	C.I値
好き	If[造形が有機的]Then[選考は好き]	2/3
	If[カラーが色彩系]and[イメージがパーソナル]Then[選考は好き]	1/3
	If[ドアタイプが2ドア]and[イメージがパーソナル]Then[選考は好き]	1/3
どちらでもない	If[カラーが白黒系]and[造形が曲線的]Then[選考はどちらでもない]	2/3
	If[造形が曲線的]and[ドアタイプが4ドア]Then[選考はどちらでもない]	2/3
	If[イメージがフォーマル]Then[選考はどちらでもない]	1/3
	If[造形が曲線的]and[イメージがパーソナル]Then[選考はどちらでもない]	1/3

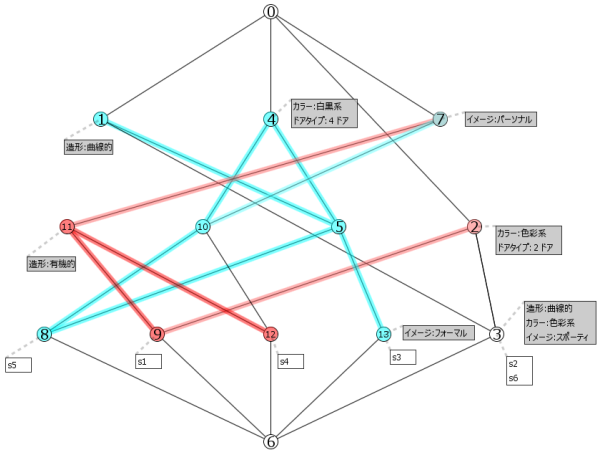

図 3:決定ルール投影ハッセ図(自動車の選好データ)

図における赤く着色されたノードとリンクは決定クラス「好き」,青く着色されたノードとリンクは決定クラス「どちらでもない」の決定ルールをあらわしている.システムを用いた分析結果から解釈した情報を以下に示す.

(1) 複数の決定クラスに属する属性値の判別

図 3 における 7 番のノード(属性値「イメージがパーソナル」)に着目する.このノードは赤と青両方の色で着色されており,下位ノードに赤いソード(11 番ノード)と青いソード(10 番ノード)を 1 つずつ持っている.このことから,属性値「イメージがパーソナル」は決定クラス「好き」と「どちらでもない」両方の決定ルールに含まれていることが分かる.すなわち,属性値「イメージがパーソナル」は,単体だとどちらの決定クラスの確定にも影響せず,他の属性値(「造形が有機的」など)と組みあわさることではじめて「好き」あるいは「どちらでもない」の決定ルールとなり,決定クラスの確定に影響するようになるといえる.

次に,表 4 より,決定クラス「好き」と決定クラス「どちらでもない」それぞれの決定ルール中に含まれる属性値に着目すると,属性値「イメージがパーソナル」は両方の決定ルールに含まれていることが分かる.また,属性値「イメージがパーソナル」が含まれる決定ルールに着目すると,属性値「イメージがパーソナル」は,単体ではなく,必ず何らかの他の属性値と組みあわさることで決定ルールを構成していることがわかる.これは先述した図 3 から読み取れる情報と同様だが,ラフ集合解析の結果のみを用いてこのような属性値を見つけるには,複数の決定クラスに共通して含まれている属性を探し,またその属性が他の属性とどのように組み合わさり決定ルールを構成しているのかを調べる必要があり,見落とす可能性も高い.

このことから,図 3 のような決定ルール投影ハッセ図を使用すると,ノードの色およびそのノードの下位ノードの色によって,こういった複数の決定クラスに属する特殊な属性値を従来のラフ集合解析と比べて容易に判別することができるといえる.

(2) 決定ルールを確定させる要因の大小の判別

図 3 における 12 番ノード(サンプル「s4」)に着目する.このノードは赤く着色されていることから,サンプル s4 は決定クラス「好き」に属することが分かる.しかし,12 番ノードの上位ノードに着目すると,12 番ノードは決定クラス「それほど」に属する 10 番ノードに包含されていることが分かる.これより,サンプル s4 は決定クラス「好き」に属していながら,決定クラス「それほど」の原因となりうる属性値を持っており,選好が「どちらでもない」に 変わる可能性を持っているといえる.

次に,表 3 における決定クラス「好き」の下近似であるサンプル s1 と s4 に着目する.表 3 における s1 の属性値と表 4 の決定表より,s1 は表 4 に示す決定クラス「好き」の決定ルール条件部[造形が有機的] [カラーが色彩系] [イメージがパーソナル] [ドアタイプが 2 ドア] のすべてに当てはまっている.一方,s4 は決定クラス「好き」の決定ルール条件部のうち [造形が有機的]の 1 つしか当てはまっていない.このことから,サンプル s4 はサンプル s1 に比べ,決定クラス「好き」を確定させる要因が小さいといえる.また,s4 が持っている属性値「ドアタイプが4ドア」と「カラーが白黒系」は表 4 において決定クラス「どちらでもない」の決定ルール中にのみ含まれる属性値であり,この属性値を持っているサンプル s4 は決定クラスが「どちらでもない」になる要因が s1 に比べて大きく,s4 は s1 よりも決定クラス「どちらでもない」に近いサンプルであるといえる.これは先述した図 3 から読み取れる情報と同様だが,ラフ集合解析の結果のみを用いてこのような属性値を見つけるには,サンプルごとに決定ルールをどれだけ満たしているかを 1 つずつ調べる必要がある.

このことから,図 3 のような決定ルール投影ハッセ図を使用すると,ノードの色およびそのノードの上位ノードの色によって,各サンプルの決定クラスを確定させる要因の大小を,従来のラフ集合解析と比べて容易に、また短時間で判別することができるといえる.

(3) 属性値どうしの包含関係の判別

図 3 における9番ノードと 12 番ノードおよびそれらの上位ノードである 11 番ノードより,決定クラス「好き」の下近似であるサンプルは属性値「造形が有機的」を有しているという共通点がある.これより,属性値「造形が有機的」は決定クラス「好き」の確定において,大きな要因となっているといえる.ここで図 3 をみると,属性値「造形が有機的」に対応する 11 番ノードは,上位ノードに 7 番ノード(属性値「イメージがパーソナル」)を持つ.よって,属性値「イメージがパーソナル」は属性値「造形が有機的」を包含する属性値であり,同様に決定クラス「好き」の確定において,大きな要因となっている可能性があるといえる.

一方,表 4 の決定ルールのみを用いると If[造形が有機的]Then[選好は好き] の決定ルールより,決定クラス「好き」の下近似であるサンプルが属性値「造形が有機的」を有していることは読み取れるが,属性値「イメージがパーソナル」が属性値「造形が有機的」を包含しており,決定クラス「好き」の確定に大きく影響している可能性があることは読みとることができない.

このことから,図 3 のような決定ルール投影ハッセ図を使用すると,従来のラフ集合解析の結果から得られる重要度の高い属性値に加え,決定ルールのみからでは読みとることができない,その属性値と包含,被包含関係にある属性値を見つけ出すことができるといえる.

4.3 スナック菓子の選好データを用いた比較

2 つめの検証用データとして，表 5 のスナック菓子の選好データをあげる．これは，福井らの研究[3]において使用された決定表である．従来のラフ集合解析によって得られる決定ルールを表 6 に，本研究で開発したシステムから得られる決定ルール投影ハッセ図を図 4 に示す．決定属性は「選好」とした．

表 5:決定表(スナック菓子の選好データ)

サンプル	原料	味	形	口当たり	選好
s1	小麦	塩	丸	堅め	それほど
s2	じゃが	しょうゆ	丸	堅め	好き
s3	もろこし	塩	丸	ソフト	それほど
s4	小麦	しょうゆ	四角	ソフト	好き
s5	じゃが	塩	四角	堅め	それほど
s6	じゃが	しょうゆ	丸	ソフト	好き
s7	もろこし	塩	丸	ソフト	好き

表 6:決定ルール(スナック菓子の選好データ)

決定クラス	決定ルール	C.I値
好き	If[味がしょうゆ] Then[選考は好き]	3/4
	If[原料がじゃが]and[形が丸]Then[選考は好き]	2/4
	If[原料が小麦]and[形が四角]Then[選考は好き]	1/4
	If[形が四角]and[口当たりがソフト]Then[選考は好き]	1/4
	If[原料が小麦]and[口当たりがソフト]Then[選考は好き]	1/4
	If[原料がじゃが]and[口当たりがソフト]Then[選考は好き]	1/4
それほど	If[味が塩]and[口当たりが堅め]Then[選考はそれほど]	2/3
	If[原料が小麦]and[味が塩]Then[選考はそれほど]	1/3
	If[原料が小麦]and[形が丸]Then[選考はそれほど]	1/3
	If[原料が小麦]and[口当たりが堅め]Then[選考はそれほど]	1/3
	If[原料がじゃが]and[味が塩]Then[選考はそれほど]	1/3
	If[味が塩]and[形が四角]Then[選考はそれほど]	1/3
	If[原料がじゃが]and[形が四角]Then[選考はそれほど]	1/3
	If[形が四角]and[口当たりが堅め]Then[選考はそれほど]	1/3

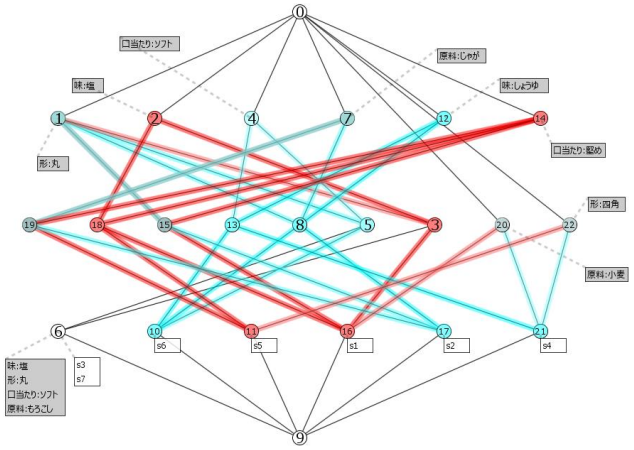

図 4:決定ルール投影ハッセ図(スナック菓子の選好データ)

図における赤く着色されたノードとリンクは決定クラス「それほど」，青く着色されたノードとリンクは決定クラス「好き」の決定ルールをあらわしている．システムを用いた分析結果から解釈した情報を以下に示す．

(1) 属性値の各決定クラスに対する影響の大小の判別

図 4 における 1 番のノード(属性値「形が丸」に対応)に着目する．このノードは下位ノードに決定クラス「好き」と「それほど」に属するノードを両方有しており，前節で述べた，複数の決定クラスに属する属性値の性質を持っている．ここで 1 番のノードの下位ノードに着目すると，赤いノードよりも青いノードの方が多い．このことから，属性値「形が丸」は「好き」と「それほど」両方の決定クラスに属しているが，どちらかといえば決定クラス「好き」の確定により大きく影響していることが読みとれる．

表 6 より属性値「形が丸」を含む決定ルールに着目すると，決定クラス「好き」における該当決定ルールの C.I 値は 2/4，決定クラス「それほど」における該当決定ルールの C.I 値は 1/3 であり，属性値「形が丸」は決定クラス「好き」に対する影響の方が大きいことが分かる．これは先述した図 4 から読み取れる情報と同様だが，ラフ集合解析の結果のみを用いて各決定ルールへの属性値の影響を比較する場合は，対象の属性値が含まれる決定ルールの C.I 値を調べる必要がある．

このことから，図 4 のような決定ルール投影ハッセ図を使用すると，ノードの色およびそのノードの下位ノードの色によって，各属性値がどの決定ルールにどれほど影響しているかを，従来のラフ集合解析と比べて容易に判別することができるといえる．

(2) 決定クラスの確定に影響を与えない属性値の判別

図 4 における 6 番のノード(属性値「原料がもろこし」に対応)は着色されていない．これはどの決定ルールにも含まれていない要素であることを示している．そのため，もろこしは選好にほとんど影響を及ぼさない原料であるといえる．

表 6 の各決定ルールに着目すると，すべての決定ルール中に属性値「原料がもろこし」が含まれていないことが分かる．よって，「原料がもろこし」はすべての決定クラスの確定にまったく関与しない属性値であることが分かる．これは先述した図 4 から読み取れる情報と同様だが，ラフ集合解析の結果のみを用いてこのような属性値を見つけるには，すべての決定ルール中に一度もあらわれていない属性値を調べる必要がある．

このことから，図 4 のような決定ルール投影ハッセ図を使用すると，ノードの色によって，決定クラスの確定に影響を与えない属性値を，従来のラフ集合解析と比べて容易に判別することができるといえる．

5. まとめ

評価実験の結果，前述のように本システムの分析結果からしか得られない概念構造を得ることができた．また，今後の課題を以下にあげる．

1) 入力される決定表の属性数やサンプル数が多くなると処理速度が極めて遅くなるため，形式概念分析アルゴリズムがより高速に動作するよう改善する必要がある．
2) 現在のシステムはすべての決定ルールを同一のハッセ図上に描画するため，決定ルール数が多い決定表を分析すると，分析結果がきわめて複雑となってしまう．そのため，各決定ルールをハッセ図上で正確に判別できるようなインターフェースに改良する必要がある．
3) より高度なデータマイニングができるよう，ラフ集合を用いた縮約を応用したハッセ図の簡単化機能や，決定ルールにもとづくハッセ図の分解機能などを実装する必要がある．

参考文献

[1] B.A.Davey，H.A.Priestley：Introduction To Lattices And Order，Cambridge University Press，p.p.65-65，2002.
[2] 井上勝雄，原田利宣，椎塚久雄，工藤康生，関口彰：ラフ集合の感性工学への応用，海文堂，p.p.213-213，2009.
[3] 福井正康，石丸敬二，尾崎誠：社会システム分析のための統合化プログラム 18 ーラフ集合分析・不確実性分析ー，福山平成大学経営研究，No.8，p.p.89-107，2012.