1K2-5

形式概念間の距離を用いた概念束の簡素化の評価

Evaluation of Concept Lattice Reduction Using Distance between Formal Concepts

石樽 隼人 *1 Hayato Ishigure　寺町 太貴 *2 Taiki Teramachi　武藤 敦子 *1 Atsuko Mutoh　森山 甲一 *1 Koichi Moriyama　犬塚 信博 *1 Nobuhiro Inuzuka

*1名古屋工業大学 大学院工学研究科情報工学専攻
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

*2名古屋工業大学 工学部情報工学科
Dept. of Computer Science, Faculty of Engineering, Nagoya Institute of Technology

Formal Concept Analysis (FCA) is a data analysis method using formal concepts and a concept lattice. Formal concepts are concepts defined mathematically and a concept lattice is the structure of formal concepts. It is a problem of FCA that the concept lattice size increases substantially and the lattice becomes more complex as the data size increases. Thus various methods for reducing a concept lattice have been proposed. In this paper, we defined the correspondence between a formal concept in the original concept lattice and one in the reduced concept lattice. In addition, we evaluated reduction methods using the correspondence.

1. はじめに

形式概念分析は，束論の応用として提案されたデータ分析手法である [Wille 82]．形式概念分析では，対象と属性の対応関係を表す形式文脈から，数学的に定義された概念である形式概念を得，得られた形式概念を利用してデータ分析を行う．また，形式概念がなす構造である概念束の可視化により，データの構造の理解を助けることができる．

一方で，データの増大に従い，概念束が急速に大きく複雑になることが，形式概念分析の問題として知られている．そのため，概念束の簡素化により，サイズを小さくし，構造を単純化する必要がある．これまでに，概念束の簡素化手法として，氷山概念束 [Stumme 01]，特異値分解を用いた手法 [Gajdoš 04]，安定度 [Kuznetsov 07]，属性推定を用いた手法 [Ishigure 15b] など様々な手法が提案されている．

また，概念束の簡素化手法の比較評価に関する研究も行われている [Dias 15, Kuznetsov 15, 石樽 15a]．簡素化手法の評価に関する研究の一つとして，形式概念間の距離を用いた評価 [寺町 16] がある．寺町らは，距離の分布の変化を見ることで，既存の簡素化手法が，特定の望ましい性質を持っていないことを明らかにした．しかし，この研究では，定量的な分析はなされていない．本稿では，形式概念間の距離を用いて，元の概念束中の形式概念と簡素化後の概念束中の形式概念との対応関係を定義する．そして，定義した対応関係を用いて，概念束の簡素化手法を定量的に評価する．

2. 形式概念分析

2.1 形式概念

形式概念分析では，形式文脈と呼ばれるデータから，形式概念と概念束を得ることを基礎として，データ分析を行う．形式文脈 $\mathbb{K} = (G, M, I)$ は，対象集合 G，属性集合 M，その間の二項関係 $I \subseteq G \times M$ から構成される．対象 $g \in G$ と属性 $m \in M$ に対して，$(g, m) \in I$ である時，g は m を持つという．これを gIm と書く．ここで，形式文脈 $\mathbb{K}$ と $X \subseteq G$，$Y \subseteq M$ に対して，次の写像を定義する．

$$X \mapsto X^I = \{m \in M \mid gIm \text{ for all } g \in X\}$$
$$Y \mapsto Y^I = \{g \in G \mid gIm \text{ for all } m \in Y\}$$

対象集合 $A \subseteq G$ と属性集合 $B \subseteq M$ に対して，$A^I = B$ かつ $B^I = A$ である組 (A, B) を $\mathbb{K}$ の形式概念という．この時，A を外延，B を内包という．

また，二つの形式概念 $(A_1, B_1), (A_2, B_2)$ に対して，以下のように半順序が定義される．

$$(A_1, B_1) \geq (A_2, B_2) \iff A_1 \supseteq A_2 \iff B_1 \subseteq B_2$$

$\mathbb{K}$ の形式概念すべての集合とこの半順序は，束を構成する．これを概念束という．

表 1 は四つの対象，三つの属性からなる形式文脈を表す．表中の × は，対象が属性を持つことを表す．また，この形式文脈からは六つの形式概念が得られる．図 1 は表 1 の形式文脈から得られる概念束を表すハッセ図である．図中のノードは形式概念を，エッジは順序関係を表す．

表 1: 形式文脈の例

	卵生 (a)	言葉 (b)	母乳 (c)
ハト (1)	×		
ヒト (2)		×	×
カモノハシ (3)	×		×
ネコ (4)			×

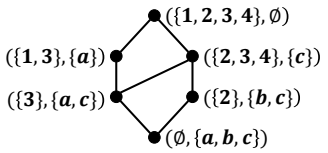


図 1: 表 1 の形式文脈から得られる概念束

連絡先: 石樽隼人，名古屋工業大学，愛知県名古屋市昭和区御器所町，h.ishigure.488@nitech.jp

2.2 形式概念間の距離

ここでは，[Blachon 07] で定義された形式概念間の距離について説明する．Blachon らは類似した形式概念のクラスタリングを行うために，形式概念間の距離を定義した．二つの形式概念 $(A_i, B_i), (A_j, B_j)$ 間の距離は，次のように定義される．

$$d_{ij} = \frac{1}{2}\frac{|A_i \triangle A_j|}{|A_i \cup A_j|} + \frac{1}{2}\frac{|B_i \triangle B_j|}{|B_i \cup B_j|}$$

ここで，$S_i \triangle S_j = (S_i \cup S_j) \setminus (S_i \cap S_j)$ である．すなわち，形式概念間の距離は，外延間の Jaccard 距離と内包間の Jaccard 距離の平均である．

3. 従来研究

3.1 概念束の簡素化

ここでは，概念束の簡素化を行う代表的な手法として，氷山概念束，特異値分解を用いた手法，安定度，属性推定を用いた手法について説明する．

氷山概念束は，データマイニングの考え方に基づく手法により簡素化された概念束である．この手法では，外延のサイズが大きな形式概念のみを抽出することで，概念束を簡素化する．得られる概念束を氷山概念束という．氷山概念束は一般には束ではなく，半束である．また，氷山概念束は元の概念束の全体ではなく，上部のみを表す．

特異値分解を用いた手法は，特異値分解による形式文脈の単純化により，概念束を簡素化する手法である．この手法では，まず形式文脈を二値行列とみなし，特異値分解を用いて低ランクでの近似を行う．ここで得られる近似行列は，実数値行列なので，直接形式文脈とみなすことができない．そのため，近似行列にしきい値を用いて二値行列に変換する．その後，得られた二値行列を形式文脈とみなして，概念束を得ることで，元の概念束を簡素化する．

安定度は，形式概念がどの程度安定しているかを表す指標である．形式概念 (A, B) の安定度は，次のように定義される．

$$\sigma(A, B) = \frac{|\{C \subseteq A \mid C^I = B\}|}{2^{|A|}}$$

安定度が大きな形式概念の内包は，少数の対象を除いても形式概念の内包となるため，ノイズに強い．安定度による簡素化では，安定度がしきい値より大きな形式概念を抽出することで，形式概念の数を減らす．

また，属性推定を用いた手法は，属性間のルールを用いて属性推定を行うことで，形式文脈の単純化を行い，概念束を簡素化する手法である．この手法では，形式文脈と概念束から，ほとんど，またはすべての対象について成立する属性間のルールを抽出する．次に抽出したルールを用いた属性推定により，ほとんどの対象について成立するルールがすべての対象について成立するように，形式文脈を更新する．属性推定により，形式文脈が単純化されるため，更新された形式文脈の概念束は，元の概念束を簡素化したものとなる．この手法の問題点として，計算量がかなり大きいことが挙げられる．

3.2 簡素化手法の評価

概念束の簡素化を行う様々な手法が提案されてきたが，簡素化手法の比較評価に関する研究は少ない．ここでは，簡素化手法の比較評価に関する研究を紹介する．

[Dias 15] では，概念束の簡素化手法が三つに分類され，それぞれどのような特徴を持つかが評価されている．[Kuznetsov 15] では，簡素化手法が [Dias 15] とは異なる形で分類されている．また，評価値に基づき形式概念を選択する手法について，ノイズ耐性などの観点で評価されている．[石樽 15a] では，形式文脈の変化や属性間の関係の維持の観点で，簡素化手法が比較評価されている．

また，概念間の距離を用いて，簡素化手法の評価を行ったものとして，[寺町 16] がある．寺町らは，概念束の簡素化手法が持つべき望ましい性質について考察している．そして，次の二つの観点に関して，簡素化手法の評価を行っている．

- 簡素化の際に，形式概念の分布が維持されるか
- 簡素化の際に，形式概念の分布が一様になるか

さらに，これらの性質に対して，隣接概念間の距離の分布を比較することで，簡素化手法の評価を行っている．比較の結果，既存の手法が，形式概念の分布に関して，望ましい性質を持っていないことを指摘している．

4. 形式概念の代表関係

本稿では，元の概念束と簡素化後の概念束の分布の変化を評価するため，二つの概念束中の形式概念間の対応関係を定義する．簡素化後の概念束は，元の概念束全体を表すと仮定する．また，対応関係は，形式概念間の距離を用いて定義する．[Blachon 07] で定義された形式概念間の距離は，同一の概念束中の形式概念間に対して定義されている．しかし，本稿で扱う概念束の簡素化手法は，形式文脈上の対象集合と属性集合が変化しないため，元の概念束中の形式概念と簡素化後の概念束中の形式概念の間の距離も同様に扱うことができると考えられる．このため，本稿では [Blachon 07] で定義された距離を，二つの概念束中の形式概念間の距離を表すように拡張して扱う．形式概念の代表関係は，拡張した距離を用いて，以下のように定義する．二つの形式概念 c_1, c_2 間の距離を $distance(c_1, c_2)$ で表す。ここで次式を満たすときに，元の概念束 FC_o の形式概念 c_o は，簡素化後の概念束 FC_r の形式概念 c_r に代表される．

$$distance(c_o, c_r) = \min_{c \in FC_r} distance(c_o, c)$$

すなわち，元の概念束の形式概念は，簡素化後の概念束の中でそれに最も距離が近い形式概念に代表されると考える．また，距離が最小の形式概念が複数存在する場合には，それらすべてに代表されると考える．

5. 実験

5.1 設定

本稿では，簡素化後の概念束が元の概念束をどのように表しているかを，五つの簡素化手法に対して評価した．簡素化手法としては，氷山概念束，特異値分解を用いた手法，安定度，属性推定を用いた手法の四つと合わせて，ランダムに形式概念を選択する手法も評価した．ランダムに形式概念を選択する手法では，元の概念束に含まれる形式概念から，ランダムに特定数の形式概念を選択することで，簡素化を行う．実験では，まず元の概念束をパラメータを変えて簡素化し，元の概念束と簡素化後の概念束の代表関係を求めた．その後，代表関係をもとに，評価値を計算した．また，簡素化後の形式概念の数に基づき，各手法の評価値を比較した．非決定的な手法である属性推定を用いた手法とランダムな手法については，各パラメータごとに 100 回実験を行い，評価値の平均を計算した．

評価値は，以下の値を用いた．

- 概念束のサイズ
- 元の形式概念を代表しない形式概念の数
- 代表される元の形式概念の数の標準偏差
- 代表する簡素化後の形式概念までの距離の平均，標準偏差
- 隣接する形式概念間の距離の平均

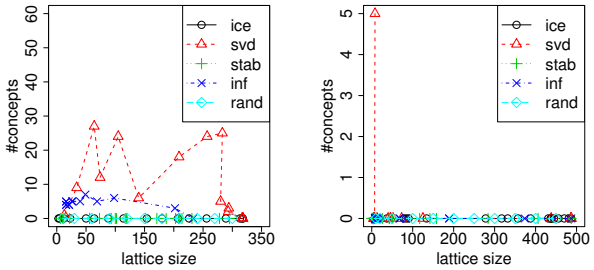
概念束のサイズは，簡素化後の概念束に含まれる形式概念の数である．元の形式概念を代表しない形式概念の数は，それが代表する元の形式概念の数が 0 となるような簡素化後の形式概念の数である．これは，元の分布と外れた位置にある形式概念の数を表す．代表される元の形式概念の数の標準偏差は，簡素化後の各形式概念が代表する元の形式概念の数の標準偏差である．この評価値が小さいほど，元の形式概念の分布を保存したまま，概念束の簡素化がされていると言える．また，代表する簡素化後の形式概念までの距離の平均，標準偏差は，簡素化前の各形式概念から，それを代表する簡素化後の形式概念までの距離の平均および標準偏差である．平均は，簡素化後の概念束が，元の概念束全体を良く表しているかを表す．標準偏差は，簡素化前の形式概念が一様な距離で表現されているかを表す．すなわち，標準偏差が小さいほど，元の形式概念の分布を一様にするように簡素化されていると言える．隣接する形式概念間の距離の平均は，簡素化後の概念束中で，順序関係において隣接する形式概念間の距離の平均である．

また，実験では二つの概念束を用いた．一つは，学生の友人関係データの概念束である．これは，学生の友人関係ネットワークから，[林 14] で提案された手法を用いて，得られた概念束である．これは，対象 395 個，属性 17 個，形式概念 318 個からなる．この概念束では，属性間に強い従属関係が存在する．すなわち，ある属性を持つならば，必ず他のある属性を持つというような関係がある．もう一つの概念束は，人工データの概念束である．これは，[Ishigure 15b] で用いられた手法により，単純な形式文脈を拡張し，ノイズを加えた形式文脈から得られた概念束である．この概念束は，対象 50 個，属性 50 個，形式概念 487 個からなる．

5.2 結果

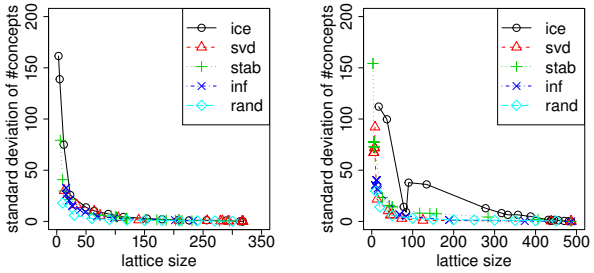
図 2 は，概念束のサイズと，元の形式概念を代表しない形式概念の数の関係を表す．図中の ice，svd，stab，inf，rand はそれぞれ氷山概念束，特異値分解を用いた手法，安定度，属性推定を用いた手法，ランダムな手法を表す．安定度などの，元の形式概念から一部を選択することで簡素化を行う手法では，元の形式概念を代表しない形式概念が存在することはない．図からも，安定度などの手法で，こうした形式概念が存在しなかったことがわかる．一方で，特異値分解や属性推定の手法では，元の形式概念を代表しない形式概念が存在した．特に，特異値分解を用いた手法と友人関係データの組み合わせでは，かなり大きな割合の形式概念が元の形式概念を代表していない場合があった．さらに詳しく調べると，元の形式概念を代表しない形式概念の多くは，属性間の従属関係に反する対象を原因とすることがわかった．また，人工データでは，元の形式概念を代表しない形式概念は，ほとんど見られなかった．

図 3 は，概念束のサイズと，代表される元の形式概念の数の標準偏差の関係を表す．ランダムな手法は手法の性質から，元の形式概念の分布に沿って選択がされると考えられる．図からも，標準偏差が他の手法に比べ小さく，こうした性質を持つことがわかる．また，氷山概念束は，他の手法に比べて標準偏差が大きく，簡素化によって元の分布とは全く異なる形式概念



(a) 友人関係データ　(b) 人工データ

図 2: 元の形式概念を代表しない形式概念の数



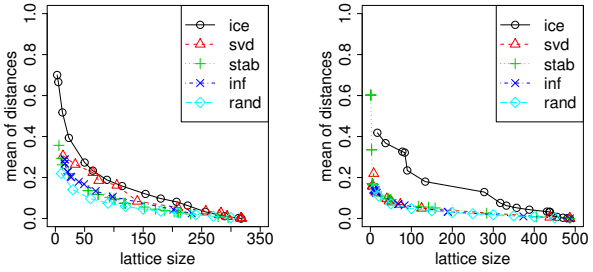
(a) 友人関係データ　(b) 人工データ

図 3: 代表される元の形式概念の数の標準偏差

の分布になることがわかる．他の手法は，友人関係データの場合に，元の形式概念の分布を保存しないことがわかる．人工データの場合には，ランダムな手法とほとんど差がなかった．人工データにおいては，複数の形式概念に代表される元の形式概念が，比較的多かったことが，この評価値に影響を与えた可能性がある．

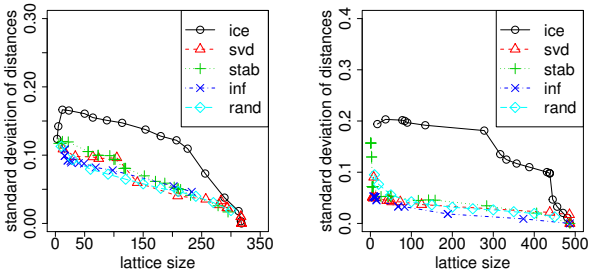
また，図 4 および 5 は，概念束のサイズと，代表する簡素化後の形式概念までの距離の平均および標準偏差の関係を表す．図 4 から，氷山概念束は，どちらのデータにおいても，元の概念束全体を表してはいないと言える．特異値分解を用いた手法では，友人関係データの場合に平均が大きくなっている部分があるが，これは元の形式概念を代表しない形式概念の存在が原因であることが推測される．また，図 5 から，氷山概念束の標準偏差が大きく，元の概念束を一様に表していないことがわかる．他の手法ではあまり差が見られなかった．また，ランダムな手法は元の概念束を一様に表す手法とは考えられないが，属性推定を用いた手法なども標準偏差が同程度なので，どの手法も一様に表しているとは考えられない．

図 6 は，概念束のサイズと，隣接する形式概念間の距離の平均を表す．図から，ランダムな手法で簡素化された概念束で，非常に距離が大きくなることがわかる．ランダムな手法では，形式概念間の順序関係が保持されないため，隣接する形式概念間の距離が大きくなることが，原因と考えられる．また，特異値分解を用いた手法と，属性推定を用いた手法では，平均距離が小さくなっている．これは，二つの手法で簡素化された概念束が，束構造を維持しているためだと考えられる．氷山概念束は，人工データの場合に，平均距離が小さかった．これは，氷山概念束が，元の概念束全体ではなく，一部の領域のみを表していることが原因だと考えられる．



(a) 友人関係データ　(b) 人工データ

図 4: 代表する簡素化後の形式概念までの距離の平均



(a) 友人関係データ　(b) 人工データ

図 5: 代表する簡素化後の形式概念までの距離の標準偏差

6. まとめと今後の課題

本研究では，概念束の簡素化手法を評価するため，形式概念間の距離を用いて，形式概念間の代表関係を定義した．また，代表関係を用いて，概念束の簡素化手法を定量的に評価した．実験結果から，既存の簡素化手法が，元の形式概念間の分布を維持する，または分布を一様化するという性質を持っていないことを再確認した．ランダムな手法では，分布を維持すると考えられるが，この手法は形式概念間の順序関係を維持しないことを示唆する結果も得られた．また，特異値分解を用いた手法や属性推定を用いた手法で，元の形式概念を代表していない形式概念が多く現れる場合があることがわかった．これらの手法によって簡素化した概念束から，元の概念束の性質を観察する際には注意が必要となる．

本稿では，形式概念の分布全体を一括して，一つの指標として調べたが，手法の差を明らかにするために，分布を局所的に観察する必要があるかもしれない．また，簡素化後の概念束が，元の形式概念間の関係を維持しているかも評価したい．形式概念の分布に関して望ましい性質を持つ簡素化手法の調査も，今後の課題として挙げられる．

参考文献

[Blachon 07] Blachon, S., Pensa, R. G., Besson, J., Robardet, C., Boulicaut, J. F., and Gandrillon, O.: Clustering Formal Concepts to Discover Biologically Relevant Knowledge from Gene Expression Data, *In Silico Biology*, Vol. 7, No. 4, 5, pp. 467–483 (2007)

[Dias 15] Dias, S. M. and Vieira, N. J.: Concept lattices reduction: Definition, analysis and classification, *Expert Systems with Applications*, Vol. 42, No. 20, pp. 7084–7097 (2015)

[Gajdoš 04] Gajdoš, P., Moravec, P., and Snášel, V.: Concept Lattice Generation by Singular Value Decomposition, in *Proceedings of CLA 2004* (2004)

[Ganter 99] Ganter, B. and Wille, R.: *Formal Concept Analysis: Mathematical Foundations*, Springer (1999)

[林 14] 林 宏紀, 伊東 樹希, 西尾 典晃, 武藤 敦子, 犬塚 信博：エゴセントリックネットワークと形式概念分析を利用した社会ネットワーク分析, 人工知能学会論文誌, Vol. 29, No. 1, pp. 177–181 (2014)

[石樽 15a] 石樽 隼人, 武藤 敦子, 松井 藤五郎, 犬塚 信博：形式概念束の簡素化手法の評価, 平成 27 年度電気・電子・情報関係学会東海支部連合大会講演論文集 (2015)

[Ishigure 15b] Ishigure, H., Mutoh, A., Matsui, T., and Inuzuka, N.: Concept Lattice Reduction Using Attribute Inference, in *Proceedings of IEEE GCCE 2015*, pp. 108–111 (2015)

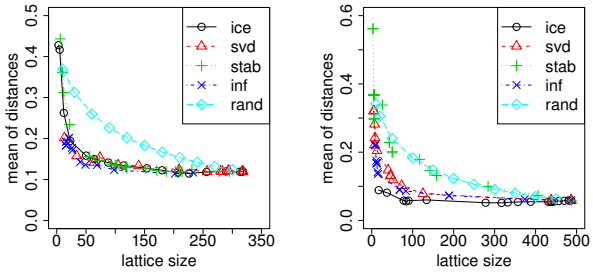
[Kuznetsov 07] Kuznetsov, S., Obiedkov, S., and Roth, C.: Reducing the Representation Complexity of Lattice-Based Taxonomies, in *Proceedings of ICCS 2007*, pp. 241–254, Springer (2007)

[Kuznetsov 15] Kuznetsov, S. O. and Makhalova, T. P.: Concept interestingness measures: a comparative study, in *Proceedings of CLA 2015*, pp. 59–72 (2015)

[Stumme 01] Stumme, G., Taouil, R., Bastide, Y., and Lakhal, L.: Conceptual Clustering with Iceberg Concept Lattices, in *Proceedings of FGML 2001* (2001)

[寺町 16] 寺町 太貴, 石樽 隼人, 武藤 敦子, 森山 甲一, 犬塚 信博：形式概念束の簡約化のための概念間の距離に関する検討, 情報処理学会第 78 回全国大会講演論文集 (2016)

[Wille 82] Wille, R.: Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts, in *Ordered Sets*, pp. 445–470, Springer (1982)



(a) 友人関係データ　(b) 人工データ

図 6: 隣接する形式概念間の距離の平均