形式概念束の簡約化法評価のための概念間距離に関する検討

犬塚・武藤研究室　知能系

No. 21115097　寺町 太貴

1　はじめに

数学的に定義された概念に基づき、データ分析を行う手法として形式概念分析 [1] がある。形式概念分析では、概念の構造を表す概念束を可視化することで、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のサイズが急速に大きくなると、構造の理解が難しくなるため、概念束の簡約化が必要になる。簡約化の手法にはいくつか種類がありそれぞれに特徴があるが、簡約化の際に概念束間の距離が偏るなど適切に行われていない場合がある。

本研究は簡約化の一つの尺度として概念間の距離を定義し、簡約化前後の距離について検討した。

2　形式概念と概念束

形式概念分析では対象の集合 G と属性の集合 M、G と M の間の二項関係 $I \subseteq G \times M$ を扱う。三つ組 $\mathbb{K} = (G, M, I)$ を形式文脈という。また gIm である時、「対象 g は属性 m を持つ」という。ここで二つの集合 $X \subseteq G$ と $Y \subseteq M$ について、X のすべての対象が共通して持つ属性の集合が Y であり、Y のすべての属性を持つ対象の集合が X である時、組 (X, Y) を形式概念という。また、対象集合の包含関係により、形式概念の順序が定義される。これを概念束という。

3　形式概念における距離の分布

隣接する 2 つの概念 $(A_1, B_1), (A_2, B_2)$ 間の距離 d_{12} を以下の数式 [2] で定義する。

$$d_{12} = 1 - \left\{ \frac{1}{2}\frac{|A_1 \cap A_2|}{|A_1 \cup A_2|} + \frac{1}{2}\frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} \right\}$$

簡約化に必要な性質は、目的に応じていくつか存在し、形式概念の数を単純に減少させる、形式概念の性質を維持しつつ個数を減少させる、元の形式概念の代表的なものを選んでいる、アルゴリズムの計算量等が挙げられる。本研究では主に、簡約化前後の距離の分布の変化に注目する。本稿では概念束の簡約化として氷山概念束、特異値分解を用いた手法、安定度による手法、属性推定を用いた手法を取り上げ、簡約化による距離の分布の変化を調べる。

4　実験

平成 24・25 年度入学の名古屋工業大学情報工学科の学生の友人関係に基づく概念束と、人工的な概念束にノイズをのせたデータの二つを利用し、各データに対して 4 つの手法を用いて簡約化を行い、隣接する概念間の距離の分布を比較した。4 つの簡約化についてこの性質を調べ、各手法に対して最大・最小・平均に関してそれぞれ比較した。図 1 は簡約化前の友人関係データの概念間の距離の最大値の分布で、図 2 は友人関係データに対して各手法を用いて簡約化後の各概念間の距離の最大値の分布をとったものである。

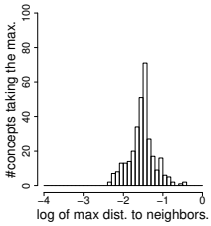
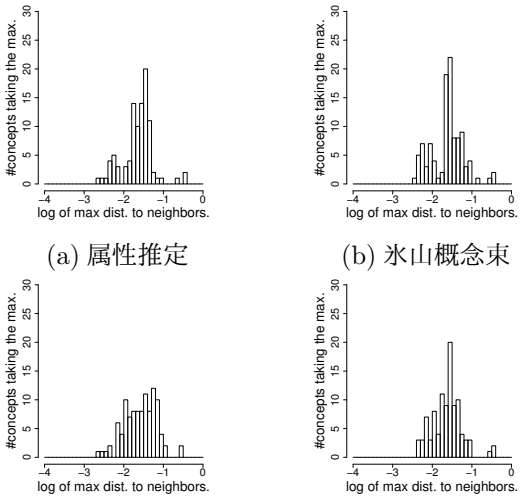


図 1: 友人関係データ簡約前の隣接概念への最大距離



(a) 属性推定　(b) 氷山概念束

(c) 特異値分解　(d) 安定度

図 2: 友人関係データ簡約後の隣接概念への最大距離

属性推定を用いた手法では、分布の形を比較的維持している一方、概念間距離の均一化という面では有効な結果が出ていない。特異値分解を用いた手法では、分布の維持は比較的できているが均一化の面では属性推定を用いた手法よりも、傾向が弱い結果となった。安定度を用いた手法では分布を維持する傾向は低いものの、均一化の傾向が他の手法に比べて強いと言うことが観察できた。氷山概念束では、分布の維持・概念間の距離の均一化両方に対して強い傾向を見つけることができなかった。

5　まとめ

概念束簡約化に求める性質として隣接概念への距離分布に関する性質を提案した。その中で、特異値分解を用いた手法と属性推定を用いた手法は比較的簡約化前の距離の分布を維持する傾向にあり、安定度による手法は分布を均一方向に修正する傾向が見られた。

参考文献

[1] Ganter, B. and Wille, R.: Formal Concept Analysis: Mathematical Foundations, Springer, 1998.

[2] Blachon, S., *et al.*: Clustering Formal Concepts to Discover Biologically Relevant Knowledge from Gene Expression Data, In Silico Biology **7**, 4–5, pp. 467–483, 2007.