3B1-04

疑似クリークに基づく近似的形式概念の抽出

Extracting Approximate Formal Concepts Based on Pseudo-Cliques

大久保 好章　　原口 誠
Yoshiaki OKUBO　Makoto HARAGUCHI

北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Division of Computer Science, Graduate School of Information Science and Technology, Hokkaido University

We discuss in this paper an approximation of *Formal Concepts*. A formal concept is precisely defined as a pair of an extent O (a set of objects) and an intent A (a set of attributes), where O and A are necessary and sufficient for identifying each other. For several formal concepts, if their extents and intents considerably overlap, it would be reasonable to combine them into an approximate formal concept still preserving almost original characteristics. We propose a definition of such an approximate formal concept. Its completeness and weak-soundness are also presented based on two approximation parameters α and β.

1. はじめに

著者等はこれまで，文献 [原口 02] をその始まりとして，クリーク探索に基づくクラスタ抽出の研究を行なってきた (例えば [Haraguchi 06a])．対象とするデータ (個体) 群を，所与の類似関係のもとで無向グラフ表現し，その極大クリークを探索することで，クラスタ抽出を行なう．特に，評価値が上位 N (Top-N) のクラスタのみを，分枝限定を利用して効率良くピンポイントに抽出することが大きな特徴である．

クラスタ抽出においては，クラスタの**解釈・意味付け**が重要であるが，文献 [Haraguchi 06b] では，こうした点を，**形式概念解析** (*Formal Concept Analysis*) [Ganter 99] の枠組で議論している．形式概念解析では，**外延** (*Extent*) と**内包** (*Intent*) の組として概念を定義し，これを形式概念と呼ぶ．抽出すべきクラスタを形式概念に限定することで，クラスタ (外延) の解釈を，その内包の言葉で明確に語ることが可能となる．

文献 [Haraguchi 06b] では，クラスタ抽出問題を，内包に関する制約を満たし，かつ，外延の評価値が Top-N の形式概念を求める問題として定式化し，その計算アルゴリズムを設計・実装した．それは著者らの従来アルゴリズムの拡張であり，そこでも分岐限定を利用した効率良い探索が可能である．この様な内包制約のもとでの Top-N 形式概念は，データマイニングにおける**飽和集合**の高速列挙アルゴリズム (例えば *LCM* [Uno 04]) の出力を後処理することでも抽出可能である．しかし，内包に関する制約が厳しい場合には，列挙された膨大数の飽和集合の中から条件を満たすものを選ぶ必要があり，計算時間が増大する [Haraguchi 06b]．

この様に，評価値が Top-N であるものに抽出対象を限定することで，効率良い探索が可能となる一方，形式概念に基づくクラスタ抽出においても，これまでの研究で観察された**クラスタの重複問題** [Okubo 05]，すなわち，重複の大きなクラスタが，Top-N の大部分を占め，得られたクラスタ間に大きな違いを見出せない状況が予想される．文献 [Okubo 05] では，**疑似クリーク**の考えを導入することでこうした問題の緩和を試みた．これまでの疑似クリークの考え方にならい，本稿でも，重複の度合が大きな複数の形式概念をひとつの近似的な形式概念とみなすことで，形式概念における重複問題の緩和を試みる．

連絡先: 大久保 好章・原口 誠
北海道大学大学院情報科学研究科コンピュータサイエンス専攻
〒 060-0814 札幌市北区北 14 条西 9 丁目
TEL：011-706-7161
E-mail：{ yoshiaki, mh }@ist.hokudai.ac.jp

2. 準備

$\mathcal{O}$ を**個体**の集合, $\mathcal{A}$ を**属性**の集合とし，各個体 $o \in \mathcal{O}$ を，それが有する属性の集合 $A_o \subseteq \mathcal{A}$ で表す．ここで，$\mathcal{O}$ と $\mathcal{A}$ の組 $<\mathcal{O}, \mathcal{A}>$ を**文脈**と呼ぶ．

文脈 $<\mathcal{O}, \mathcal{A}>$ に関して，写像 $\varphi : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{A}}$ および $\psi : 2^{\mathcal{A}} \rightarrow 2^{\mathcal{O}}$ を考える．ここで，個体集合 $O \subseteq \mathcal{O}$ と属性集合 $A \subseteq \mathcal{A}$ について，

$$\varphi(O) = \{a \in \mathcal{A} \mid \forall o \in O \;\; a \in A_o\} = \bigcap_{o \in O} A_o,$$

$$\psi(A) = \{o \in \mathcal{O} \mid A \subseteq A_o\}$$

とする．つまり，φ は O 中のすべての個体が共有する属性の集合を，一方，ψ は A 中のすべての属性を有する個体の集合を返す写像である．

これら写像のもと，個体集合 $O \subseteq \mathcal{O}$ と属性集合 $A \subseteq \mathcal{A}$ について，$\varphi(O) = A$ かつ $\psi(A) = O$ が成り立つ時，O と A の組 $FC = (O, A)$ を**形式概念** (*Formal Concept*) と定める．ここで，O と A をそれぞれ FC の**外延** (*extent*)，および，**内包** (*intent*) と呼ぶ．φ と ψ の定義より，$\psi(\varphi(O)) = O$ かつ $\varphi(\psi(A)) = A$ であることは明らかである．すなわち，形式概念とは，写像 φ と ψ に関して**閉じた** (*closed*) 個体集合 O と属性集合 A の組で与えられる．O は，A 中のすべての属性を有する個体のみから成り，かつ，それら以外にこうした個体は存在しない．同様に，A は，O 中のすべての個体に含まれる (共有される) 属性のみから成り，かつ，それら以外にこうした属性は存在しない．以降では，**閉包** (*closure*) を生成する合成関数 $\varphi \circ \psi$ および $\psi \circ \varphi$ をどちらも *closure* で表す．

$\mathcal{O}$ および $\mathcal{A}$ を頂点集合とする 2 部グラフ $G = (\mathcal{O} \cup \mathcal{A}, E)$ を考える，ここで辺集合 E を $E = \{(o, a) \in \mathcal{O} \times \mathcal{A} \mid a \in A_o\}$ と定義する．この時，G の**極大 2 部クリーク**は，文脈 $<\mathcal{O}, \mathcal{A}>$ における形式概念と 1 対 1 に対応する．

3. 形式概念の基本性質

ここでは，形式概念に関するいくつかの理論的性質について述べる．

文脈 $\mathcal{C} =< \mathcal{O}, \mathcal{A} >$ の任意の形式概念 $FC_i = (O_i, A_i)$ および $FC_j = (O_j, A_j)$ について，

$$O_i \subseteq O_j \Leftrightarrow E_i \supseteq E_j$$

が成り立つ．

いま，形式概念間に順序関係

$$FC_i \preceq FC_j \overset{def}{\Longleftrightarrow} O_i \subseteq O_j$$

を考える．$\mathcal{C}$ における形式概念の集合を $\mathcal{FC}$ とすると，$(\mathcal{FC}, \preceq)$ は束を形成する．ここで，$FC_{i_1}, \ldots, FC_{i_k}$ の最小上界 (join) $FC_\vee$ および最大下界 (meet) $FC_\wedge$ は，それぞれ

$$FC_\vee = (closure(\bigcup_{j=1}^{k} O_{i_j}), \bigcap_{j=1}^{k} A_{i_j}),$$

$$FC_\wedge = (\bigcap_{j=1}^{k} O_{i_j}, closure(\bigcup_{j=1}^{k} A_{i_j}))$$

となる．

4. 形式概念の近似

形式概念 (O, A) は，その定義から，属性集合 A が個体集合 O を同定するために必要十分であることを意味している．この様に，形式概念は外延と内包により厳密に定義されるが，その厳密さ故に，外延と内包に関する大きな違いの無い複数の形式概念がしばしば観測される．形式概念としてのクラスタを考える際，こうした外延・内包のわずかな違いを重視して区別する立場がある一方，外延・内包がほぼ同じであることを重視して敢えてこれらを区別しない立場もある．前者に従うと，Top-N 形式概念の抽出においては，ほぼ同様の外延・内包で定義される形式概念が Top-N の大多数を占め，抽出されたクラスタ間に大きな違いを見出せない状況が起こり得る．ここでは後者の立場に立ち，外延・内包がほぼ同じである形式概念を，ひとつの近似的な形式概念にまとめあげることで，文脈 (データ) 中の概念的まとまりをより大まかに捉えることを試みる．それにより，上述した状況が観測されにくくなり，より明確な違いを有するクラスタの抽出が可能となることを期待する．また，外延・内包がほぼ同じ形式概念の存在は，データ中のノイズや例外に起因することも多く，近似形式概念の導入により，これらによる影響を抑える効果も期待できる．

4.1 疑似クリークに基づく形式概念の近似

データ (個体) 群を所与の類似関係のもとでグラフ表現することで，その極大クリークとして，互いに類似した個体から成るクラスタを抽出することができる．その際，重複の大きなクリーク，つまり，ほぼ同様の個体から成るクラスタがしばしば抽出されるため，文献 [Okubo 05] では，これらを統合してひとつのクラスタと見做した．統合後はもはや厳密な意味でのクリークを成さないため，これを**疑似クリーク**と定義した．特に，統合前のそれぞれのクリークの重複部分は，それらを統合する根拠となる重要な部分である．よって，これを疑似クリークの**核**と呼び，それ以外の部分とは明確に区別する．形式概念の近似もこの考えに基づいて考察する．

いま，外延・内包がほぼ同じ形式概念 $FC_i = (O_i, A_i)$ と $FC_j = (O_j, A_j)$ について，それぞれから定まる以下の個体集合

$$O = O_i \cup O_j, \quad O' = O_i \cap O_j,$$

と，属性集合

$$A = A_i \cup A_j, \quad A' = A_i \cap A_j$$

を考える．FC_i と FC_j の外延・内包に大きな違いがない場合，O と O'，および，A と A' も大きく違わないことが期待できる．別の言い方をすると，FC_i と FC_j の両者を統合して得られる O と A は，FC_i と FC_j に共有される O' と A' [*1] をそれぞれ近似的に表現していると言えよう．このことから，(O, A) は FC_i と FC_j の両者を意味する近似的な形式概念であり，その主要な個体集合と属性集合は O', A' であると考えることは極自然であろう．先に述べた通り，形式概念は極大 2 部クリークとの対応がとれ，この統合操作は，疑似的な 2 部クリークを考えることに相当している．

以下では，こうした近似形式概念をより正確に定義する．

4.2 (α, β)-近似形式概念

これまでの議論で，外延・内包がほぼ同じ形式概念 $FC_i = (O_i, A_i)$ と $FC_j = (O_j, A_j)$ に対して，$(O = O_i \cup O_j, A = A_i \cup A_j)$ なる近似形式概念を対応させることを述べた．ここで，形式概念の性質より，FC_i と FC_j の最小上界 $FC_\vee$，および，最大下界 $FC_\wedge$ はそれぞれ

$$FC_\vee = (closure(O_i \cup O_j), A_i \cap A_j)$$

$$FC_\wedge = (O_i \cap O_j, closure(A_i \cup A_j))$$

となり，特に，

$$closure(O_i \cup O_j) \supseteq O_i \cup O_j$$

$$closure(A_i \cup A_j) \supseteq A_i \cup A_j$$

である．よって，$closure(O_i \cup O_j)$ と $O_i \cap O_j$，および，$closure(A_i \cup A_j)$ と $A_i \cap A_j$ との差が小さいことを要請すれば，必然的に $O_i \cup O_j$ と $O_i \cap O_j$ および $A_i \cup A_j$ と $A_i \cap A_j$ の差も小さくなり，結果として，FC_i と FC_j の外延・内包がそれぞれほぼ同じものとなり，先の議論に従って両者を統合できることになる．より一般的には，$FC \preceq FC'$ なる，形式概念 $FC = (O, A)$ と $FC' = (O', A')$ について，それぞれの外延・内包の差が小さいことを要請すれば，$FC \preceq FC'' \preceq FC'$ なる任意の形式概念 FC'' の外延と内包はそれぞれほぼ同じものとなり，それらを統合できることになる．

これをもとに近似形式概念を以下の通り定義する．

定義：(α, β)-**近似形式概念**

文脈 $\mathcal{C} =< \mathcal{O}, \mathcal{A} >$ において，$FC \preceq FC'$ なる形式概念 $FC = (O, A)$ と $FC' = (O', A')$ を考える．パラメータ α $(0 \leq \alpha < 1)$ と β $(0 \leq \beta < 1)$ に関して，

$$\frac{|O|}{|O'|} \geq 1 - \alpha \quad \text{かつ} \quad \frac{|A'|}{|A|} \geq 1 - \beta$$

である時，$\tilde{FC} = (\tilde{O} = O', \tilde{A} = A)$ を，$\mathcal{C}$ における (α, β)-**近似形式概念**と呼ぶ．ここで，O および A' の要素は，$\tilde{FC}$ の**主要な個体** (core object)，**主要な属性** (core attribute) と言われる． ■

(α, β)-近似形式概念 $\tilde{FC} = (\tilde{O}, \tilde{A})$ は以下の性質を持つ．

---

*1 O', A' は単なる共通部分であるだけでなく，FC_i と FC_j において，それらが大部分を占めることに注意．

完全性：

$\tilde{A}$ 中のすべての属性を有する個体は，必ず $\tilde{O}$ に含まれる．かつ，そうした個体は，$\tilde{O}$ 中で $(1-\alpha)$ 以上の割合を占める．

また，$\tilde{O}$ 中のすべての個体に共有される属性は，必ず $\tilde{A}$ に含まれる．かつ，そうした属性は，$\tilde{A}$ 中で $(1-\beta)$ 以上の割合を占める．

近似的健全性：

$\tilde{O}$ 中には，$\tilde{A}$ 中の一部の属性を持たない個体が存在する．しかし，その割合は高々 α であり，かつ，それらは $\tilde{A}$ 中，$(1-\beta)$ 以上の割合の属性を有する．

また，$\tilde{A}$ 中には，$\tilde{O}$ 中の一部の個体が持たない属性が存在する．しかし，その割合は高々 β であり，かつ，それらは $\tilde{O}$ 中，$(1-\alpha)$ 以上の割合の個体に有される． ■

5. 関連研究

本稿での近似形式概念は，文献 [Kanda 01] において考察された飽和集合の近似手法と密接に関連している．

トランザクションデータベース $\mathcal{TD}$ におけるアイテムの集合を $\mathcal{I}$, その部分集合で与えられるトランザクションの集合を $\mathcal{T}$ とすると，文脈 $<\mathcal{T},\mathcal{I}>$ における形式概念の内包は，$\mathcal{TD}$ における飽和集合に一致する．文献 [Kanda 01] では，飽和集合に関して，頻度のわずかな違いを無視する近似手法を提案し，その効果を確かめている．飽和集合の頻度は，形式概念の外延の大きさに対応することから，文献 [Kanda 01] では，形式概念の外延の近似を扱っていたことになる．実際，本稿での外延の近似に対する考え方はこれと同様である．

一番の違いは，形式概念の近似においては，概念的なまとまりを考慮するために，内包の近似も同時に扱う点にある．それ故に，文献 [Kanda 01] では，近似の度合を制御するパラメータがひとつであるのに対し，本稿ではふたつのパラメータを必要とする．その意味で，本稿での形式概念の近似手法は，文献 [Kanda 01] における近似手法の素直な拡張であると位置付けることができる．

6. おわりに

本稿では，Top-N 形式概念に基づくクラスタ抽出において生ずる可能性があるクラスタの重複問題に対処すべく，形式概念の近似について考察した．所与の近似パラメータの範囲で，外延・内包が同じと見做せる形式概念をひとつにまとめたものを近似的な形式概念と考える．近似形式概念は，完全かつ近似的に健全な概念であり，厳密な形式概念との誤差は，近似パラメータに基づいて正確に評価することができる．

著者らのこれまでのアルゴリズムを基礎に，現在，近似形式概念の抽出アルゴリズムの設計・実装を試みている．重複問題における近似形式概念の実際の効果については，その詳細を稿を改めて報告したい．形式概念はその厳密さ故に，データ中のノイズや例外に敏感に反応し，わずかなノイズ等の存在が重複の大きな形式概念を多数生ずることも多い．近似形式概念は，特に，こうしたノイズ等の影響を抑制するために効果的であると期待している．

参考文献

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

[Uno 04] T. Uno, M. Kiyomi and H. Arimura, LCM ver. 2: Efficient Mining Algorithm for Frequent/Closed/Maximal Itemsets, IEEE ICDM'04 Workshop FIMI'04, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-126/, 2004.

[原口 02] 原口 誠 : 最適クリーク探索に基づくデータからの概念学習，人工知能学会研究会資料，SIG-FAI-A202, pp. 63 - 66, 2002.

[Haraguchi 06a] M. Haraguchi and Y. Okubo: A Method for Pinpoint Clustering of Web Pages with Pseudo-Clique Search, Federation over the Web, International Workshop, Dagstuhl Castle, Germany, May 1 - 6, 2005, Revised Selected Papers, Lecture Notes in Artificial Intelligence, 3847, pp. 59 - 78, Springer, 2006.

[Haraguchi 06b] M. Haraguchi and Y. Okubo: An Extended Branch-and-Bound Search Algorithm for Finding Top-N Formal Concepts of Documents, Proceedings of the 4th Workshop on Learning with Logics and Logics for Learning - LLLL'06, 2006 (to appear).

[Kanda 01] K. Kanda, M. Haraguchi and Y. Okubo: Constructing Approximate Informative Basis of Association Rules, Proceedings of the 4th International Conference on Discovery Science - DS'01, Springer-LNAI 2226, pp. 141 - 154, 2001.

[Okubo 05] Y. Okubo and M. Haraguchi: Finding Significant Web Pages with Lower Ranks by Pseudo-Clique Search, Proceedings of the 8th International Conference on Discovery Science - DS'05, Springer-LNAI 3735, pp. 346 - 353, 2005.