文化系統学のための系統樹推定に関する研究

犬塚・武藤研究室　太田 章悟
学籍番号　No. 24417519

1　はじめに

生物の系統分化の歴史を明らかにする学問を生物系統学と呼ぶ。これに対し、文化の歴史的変化に関しても生物系統学の手法を用いて文化進化の系統を明らかにする学問を文化系統学 [1] と呼ぶ。従来研究では生物や文化のデータから分岐図（図 1(a)）を推定する手法（分岐図推定法）が多数提案されてきたが [2, 3]、系統樹（図 1(b)）を推定する手法は確立されたものが無い。本研究では、対象集合の分岐図から生年情報を用いて尤もらしい系統樹を推定する手法を提案する。さらに、得た系統樹が生年情報を正しく表現し、分岐図のクラスタ関係を保持していることを示し、実験によりその有用性を考察する。

2　系統樹推定

分岐図はある対象集合の進化の過程を枝分かれした図として表現したものであり、グラフとしてみたとき、対象を葉ノードに配置した木構造である（図 1(a)）。さらに図 2 に示すようにクラスタ構造を持つ。一方、系統樹はある対象がどの対象から由来したかを示す関係を表現したものであり、グラフとしてみたとき、対象を内部ノード・葉ノードともに配置している（図 1(b)）。ここで本研究では対象の**分岐図**と**生年情報**を用いて尤もらしい系統樹を推定する。推定する系統樹がそれらの情報と矛盾しないための条件を以下に提案する。

生年順条件

系統樹は対象の祖先子孫関係を表現しているため、祖先ノードの生年は子孫ノードの生年よりも早いと考えるのが妥当である。そこで、祖先ノードの生年は子孫ノードの生年よりも早いとき、その系統樹は**生年順条件**を満たすと定義する。

クラスタ条件

系統樹において、一つの部分木は一つのクラスタと考えるのは妥当である。また、形質が類似している複数の部分木とそれを束ねる親ノードについても一つのクラスタと考えることも妥当である。これらを踏まえ、クラスタのすべてのノードが、系統樹 T において以下の a),b),c) のノードからなる構造をしているとき、そのクラスタは T に**含まれる**と定義する。

a) **あるノード** R
b) R **に子ノードがあるなら** R **の 1 つ以上の子ノード**
c) *b*) **に子孫ノードがあるなら** *b*) **の子孫ノード全て**

そして、T が分岐図 C のすべてのクラスタを含んでいるとき、T は C の**クラスタ条件**を満たすと定義する。

3　提案手法

本研究では、生物や文化を対象に、以下の入出力としたアルゴリズムを提案する。

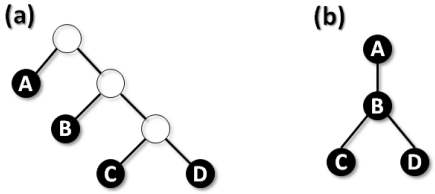

図 1: 分岐図 (a) と系統樹 (b)

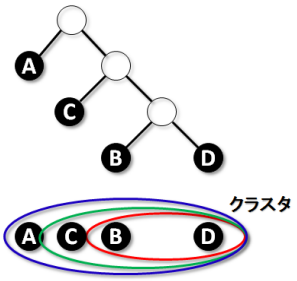

図 2: 分岐図のクラスタ構造

INPUT　C　:分岐図を表す（$N-1$）× 2 行列;
(対象数 N, 生年はアルファベット順);
OUTPUT　T　:出力する系統樹の隣接行列;

1. $T = (N \times N$ のゼロ行列）
2. $toc =$（長さ $N-1$ のベクトル）
3. **For** $i = 1$ to $(N-1)$
4. 　$a = C[i, 1]$
5. 　$b = C[i, 2]$
6. 　**If** a がクラスタ番号 **Then** $a = toc[a]$
7. 　**If** b がクラスタ番号 **Then** $b = toc[b]$
8. 　**If** a の生年が b より早い **Then** $T[a, b] = 1$
9. 　　**Else** $T[b, a] = 1$
10. 　$toc[i] =$（a, b のうち生年が早い方）
11. **Return** T

図 3: 提案手法のアルゴリズム

- 入力：全順序関係を満たす生年情報 O を持つ対象を葉ノードに配置した分岐図 C
- 出力：全順序関係を満たす生年情報 O を持つ対象を葉ノードおよび内部ノードに配置した系統樹 T

提案手法では、分岐図における内部ノードについて、その子ノードがすべて葉ノードである場合、その内部ノードを子ノードのうち生年順の最も早いノードに置き換える操作を、深さの大きい内部ノードから順に行っていくことで系統樹を得る。アルゴリズムおよび手順の例をそれぞれ図 3、図 4 に示す。

また、提案手法で得られる系統樹について、以下の 3 つの性質を満たすことを本研究の中で証明した。

定理 1　**提案手法で得られた系統樹は与えられた生年順に対し生年順条件を満たす。**

定理 2　**提案手法で得られた系統樹は与えられた分岐図に対しクラスタ条件を満たす。**

定理 3　**与えられた生年順と分岐図に対し、各条件を満たす系統樹はただ 1 つ存在する。**

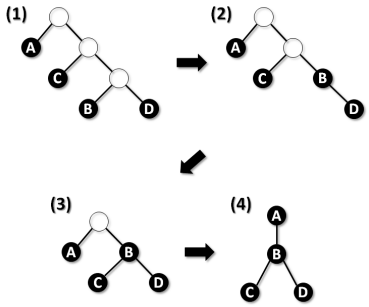

図 4: 提案手法による分岐図から系統図への変換例

4　系統モデルを用いた評価

今回の実験では提案手法の評価のため、真の系統関係が既知である対象集合として、生物や文化を模した系統モデルを用いた。これは遺伝的アルゴリズムに基づき、図 5 に示す流れで対象集合を生成するモデルである。

系統モデルによって生成した対象集合のうちいくつかをサンプリングし、それらの個体集合に分岐図推定法である UPGMA 法を適用して分岐図を推定した。得られた分岐図に提案手法を適用し系統樹を推定した後、その系統樹と真の系統関係の類似度を次の式で評価した。

$$value = \frac{|E_s \cap E_e|}{|E_s|} \qquad (1)$$

E_s と E_e は、それぞれ真の系統樹の枝集合と推定した系統樹の枝集合である。

4.1　実験結果

提案手法による結果とランダムな系統樹による結果を図 6 に示す。提案手法はランダムな系統樹よりも評価値が高くなっており、系統の復元がある程度達成できているといえる。

また、真の系統樹と推定した系統樹の一例を図 7、図 8 に示す。子孫ノードが多い影響力の強いノードがある程度再現されている (a ∼ f)。特に興味深いのが個体 e および個体 f で、真の系統樹（図 7）において、最後に系統はそれぞれ個体 e、個体 f を起点として大きく二つの系統に分かれているが、推定した系統樹（図 8）においても、同じ e と f を起点として最後に二つの系統に分かれている。

5　まとめと今後の課題

本研究では生年情報と分岐図から系統樹を推定するアルゴリズムを提案し、それにより得られる系統樹は与えた分岐図および生年情報に矛盾しないことを証明した。また、実験に適した系統モデルを提案し、それを用いた実験により推定した系統樹はランダムな系統樹と比べて大幅に尤もらしいことを示した。今後の課題としては複雑な進化現象への対応や、用いる分岐図推定法の検討などを行っていく必要があると考えられる。また、実際のデータに適用してさらなる手法の検討を進めていくことも課題に挙げられる。

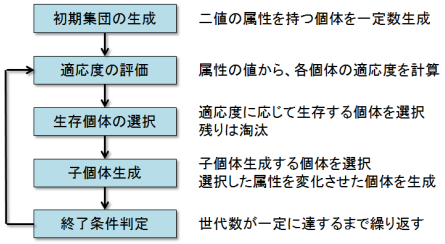

図 5: 系統モデル構成の流れ

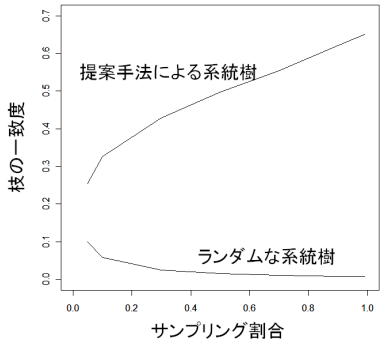

図 6: サンプリング割合と枝の一致度の変遷

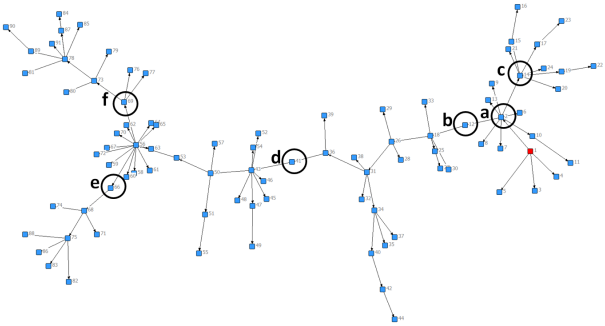

図 7: サンプリング割合が 10 %のときの、真の系統樹

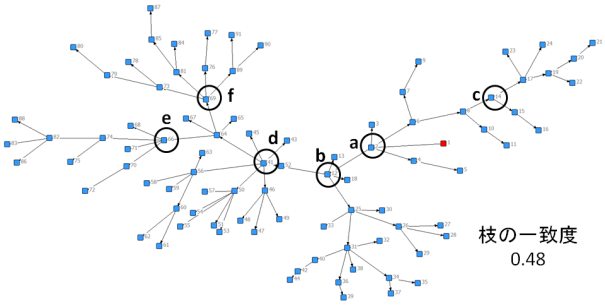

図 8: サンプリング割合が 10 %のときの提案手法による系統樹

参考文献

[1] 中尾 央, 三中 信宏: “文化系統学への招待”, 勁草書房, 2012.

[2] 斉藤 成也, 根井 正利: “The neighbor-joining method: a new method for reconstructing phylogenetic trees.”, Mol Biol Evol (1987) 4 (4): 406-25, 1987.

[3] 三中 信宏: “系統樹思考の世界”, 講談社現代新書, 2006.