

正規圧縮距離を用いたクラスタリング

久保, 浩平
九州大学理学部物理学科情報理学コース

<https://hdl.handle.net/2324/1434348>

出版情報 : COE Lecture Note. 48, pp.60-62, 2013-03-28. 九州大学マス・フォア・インダストリ研究所
バージョン :
権利関係 :





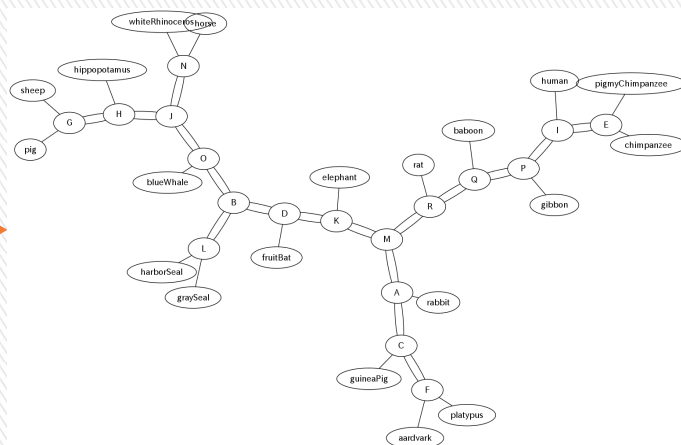
正規圧縮距離を用いた クラスタリング

九州大学 理学部 物理学科 情報理学コース
久保 浩平
2013年1月26日 13:00~14:00

	hippopotamus	Pigmy Chimpanzee	sheep	gibbon
hippopotamus	0.475728	0.964223	0.946837	0.958763
Pigmy Chimpanzee	0.964223	0.471247	0.963565	0.928932
sheep	0.946837	0.963565	0.459139	0.965563
gibbon	0.958763	0.928932	0.965563	0.468304

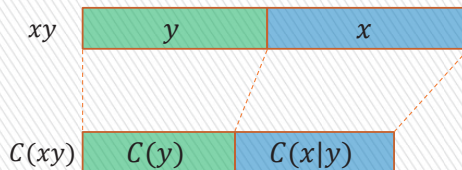
....

⋮

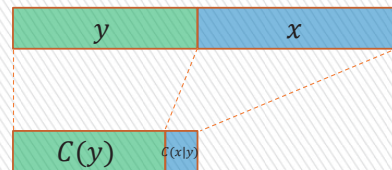


$$NCD(x, y) = \frac{\max\{C(x|y), C(y|x)\}}{\max\{C(x), C(y)\}} = \frac{C(yx) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

もしxとyが似ていなかったら...



もしxとyが似ていたら...



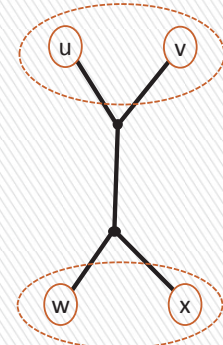
NCD



$$\text{minimize } C_T = \sum_{\{u,v,w,x\} \subseteq N} \{C_{uv|wx} : uv|wx \text{ が consistent}\}$$

> **consistent**と呼ばれる木の部分構造のコストを全て足し合わせたもの。

$$\bullet C_{uv|wx} = NCD(u, v) + NCD(w, x)$$



目的関数



フローチャート

ランダム二分木生成

$S(T)$ の値計算

以下の操作からランダムに選ぶ

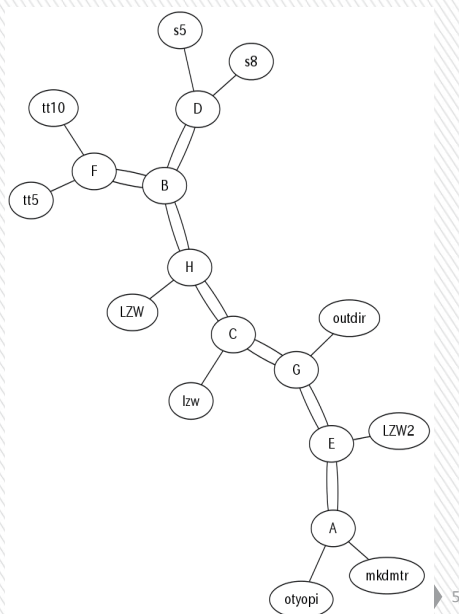
leaf_swap

subtree_swap

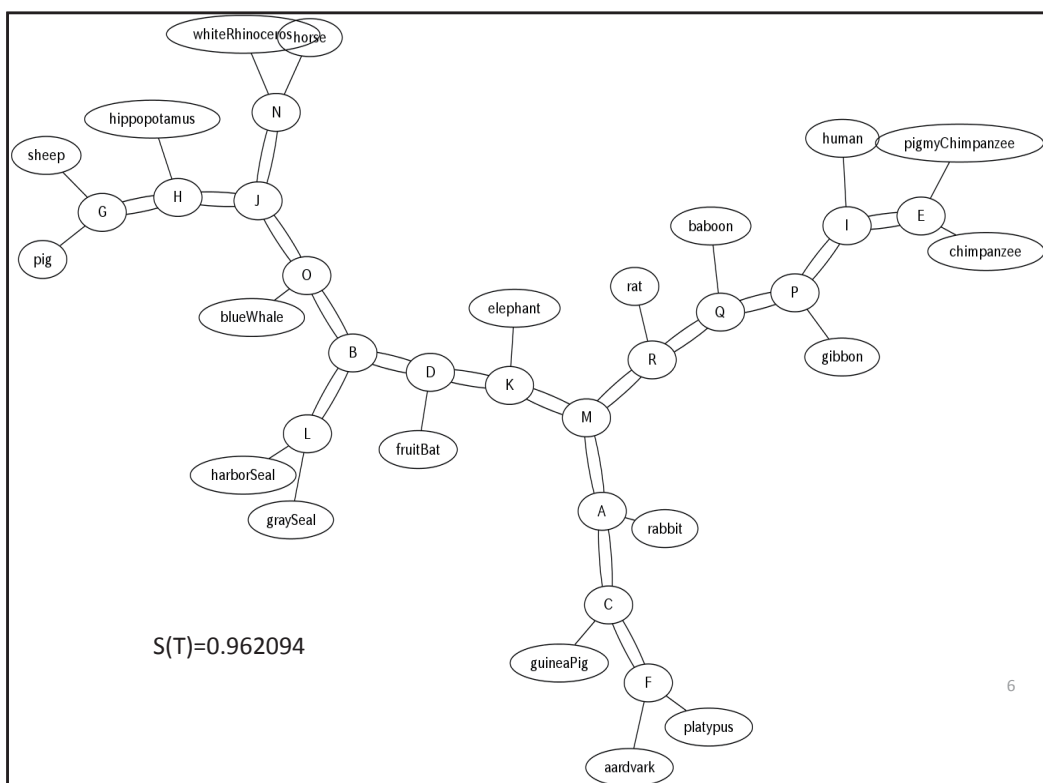
transfer

k回繰り返す

$S(T')$ の値計算



5



6