03

R-tree structure grows through stochastic process

無限に広がるビッグデータの解析

～無限階層の包含関係を持つ木構造の確率モデル～

どんな研究

本発表では無限のデータへの対策を考慮した機械学習技術を紹介します。従来のビッグデータ解析では、文字通り大きいながらも有限のサイズのデータを解析することを暗に仮定していましたが、本研究では無限のデータに対しても破綻することなく解析可能な機械学習の新しい技術を提案します。

どこが凄い

世界で初めてR木構造を無限データ解析へ応用する方法を創出しました。R木とは行列の包含関係の木構造を表すデータ構造であり、無限サイズの行列を根とするR木の全ての場合の数は無限通りとなります。これを計算機上で表現することの出来る確率モデルの構成に成功しました。

目指す未来

ビッグデータ解析はデータサイズの増加と解析アルゴリズムの更新のイタチごっこになる危険を抱え続けています。これを克服するため、あらゆるビッグデータ解析手法を無限データ解析へ発展すべく、その要素技術の拡大・確立していくことを目指しています。

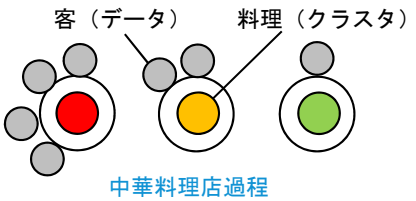
従来の無限データ解析

無限の場合の数を表現しうる確率モデル（アルゴリズム上の工夫によって無限の処理に陥らないモデル）を使い仮想的に計算機上で無限データを扱えるようにする工夫により実現。

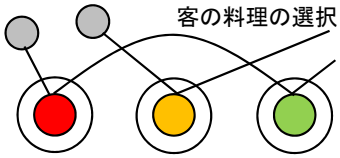
これまでに発見された無限データのための確率モデル：

- 無限のデータをクラスタリングしたい：中華料理店過程
- 無限のデータから主要構成要素を抽出したい：インド料理店過程
- 無限の行列から長方形分割模様のクラスタを抽出したい：モンドリアン過程、長方形分割過程

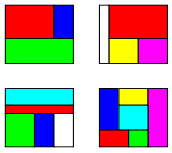
課題：無限データを扱える確率モデルは極めて少ない！

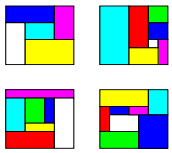


中華料理店過程



インド料理店過程

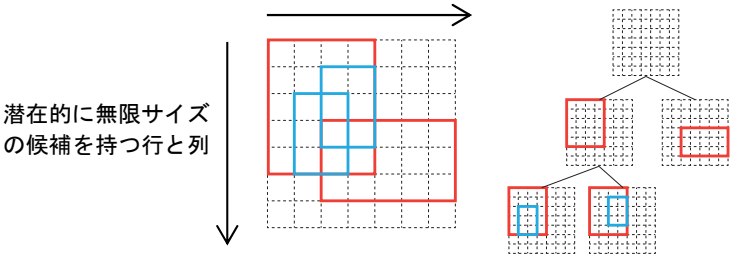

モンドリアン過程


長方形分割過程[2]

提案するモデル:R木過程

無限の行列から隠れた部分行列の無限木構造を抽出したい。

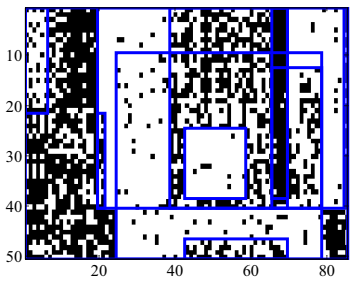
R木（各ノードに行列が対応した木構造で親ノード行列はその子ノード行列を包含するもの）と呼ばれるデータ構造に基づく、無限データのための確率モデルを初めて提案[1]。従来の無限データ解析では扱えなかった部分行列の包含関係が木構造を用いて表現可能に。



応用

関係データ解析：入力行列の行と列の並び替えとR木模様の中で、なるべく一つの領域に同じ性質のデータが集まるものを探す技術。

（例）協調フィルタリング、ネットワークの異常検知



【関連文献】

[1] 中野允裕，武小萌，森稔，木村昭悟，柏野邦夫, “R木過程,” *情報論的機械学習ワークショップ*, 2015.

[2] M. Nakano, K. Ishiguro, A. Kimura, T. Yamada, N. Ueda, “Rectangular tiling process,” in *Proc. International Conference on Machine Learning (ICML)*, 2014.

【連絡先】

中野 允裕 (Masahiro Nakano)　メディア情報研究部 メディア認識研究グループ

E-mail：nakano.masahiro(at)lab.ntt.co.jp