03 Relational data analysis by infinite trees

～ R-tree structure grows through stochastic process ～

Abstract

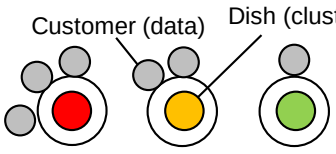
We are aiming at establishing a technical basis of infinite data analysis, which we believe is a generalized paradigm beyond big data analysis. While big data analysis has been one of the very active research topics in the machine learning field, the amount of data for the analysis has been considered finite, as implied by the term “big data” itself. However, many types of data would intrinsically grow infinitely in size, and the observed data should naturally be regarded as just a part of a potentially infinite amount of data. A key to the infinite data analysis, such as clustering and factorization, is a stochastic process, or an infinite-dimensional probabilistic model, such as Dirichlet process, Mondrian process, and rectangular tiling process [2]. Recently, we have proposed a novel stochastic process that has a capability of representing arbitrary R-trees within a matrix of infinite size. We show its principle and demonstrate its application to relational data analysis.

Infinite data analysis

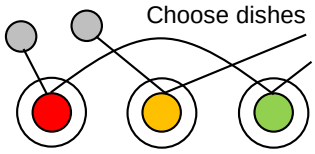
To handle infinite data on computers, infinite-dimensional probabilistic models (stochastic processes) are employed, which can represent infinitely many possible patterns.

- Clustering of infinite data：Chinese restaurant process
- Factor decomposition of infinite data：Indian buffet process
- Clustering of a matrix with infinite size：Mondrian process, and rectangular tiling process

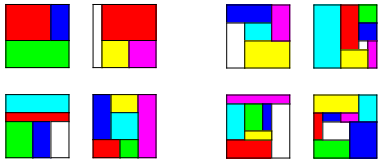
Problem: Model variations are quite limited.



Chinese restaurant process

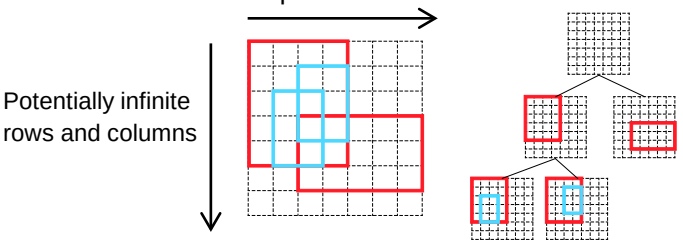


Indian buffet process


Mondrian and Rectangular tiling process

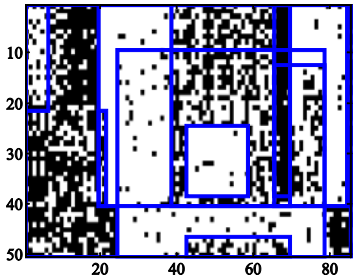
R-tree process

The objective is to extract hidden infinite R-tree structures from a matrix with infinite size. The R-tree corresponds to hierarchically structured rectangular regions within a matrix, where each node of the tree encodes a single rectangular region. This model has enabled us to represent inclusion relations embedded in a part of the infinite data matrix.



Application

Relational data analysis: finding permutations of rows and columns of a data matrix, so that similar data are clustered in the same region in a hierarchical manner.



【Reference】

[1] Masahiro Nakano, Wu Xiaomeng, Minoru Mori, Akisato Kimura, Kunio Kashino, “Stochastic process representation of R-tree,” *Information-Based Induction Science Workshop*, 2015.

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

【Contact】

Masahiro Nakano Recognition Research Group, Media Information Laboratory

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