HOKKAIDO UNIVERSITY

Title	Studies on Approximate Bayesian Computation and Speedup of Spatial Data Access Using Distributed Quadtrees [an abstract of dissertation and a summary of dissertation review]
Author(s)	Mayumbo, Nyirenda
Citation	北海道大学. 博士(情報科学) 甲第12851号
Issue Date	2017-09-25
Doc URL	http://hdl.handle.net/2115/67450
Rights(URL)	http://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	theses (doctoral - abstract and summary of review)
Additional Information	There are other files related to this item in HUSCAP. Check the above URL.
File Information	Mayumbo_Nyirenda_review.pdf (審査の要旨)

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　ニレンダ　マユンボ

審査担当者　主　査　教　授　有村 博紀
副　査　教　授　原口 誠
副　査　教　授　湊 真一
副　査　教　授　伊藤　公人　（人獣共通感染症研究センター）

学位論文題名

Studies on Approximate Bayesian Computation and Speedup of Spatial Data Access Using Distributed Quadtrees
（近似ベイズ計算と分散 4 分木を用いた空間データアクセスの高速化）

本論文では, 大規模データからの時系列予測に関して, 生物情報学におけるインフルエンザ B ウイルスの流行予測問題を考察し, その数理モデル設計と, 推論手続き設計, 高速化の側面から, データ同化に基づく流行予測システムの実現のための基礎技術について研究する.

本論文の一つ目の柱は, 互いに相互作用をもつ複数の時系列の予測のための数理モデルの設計である. 著者らは, 感染症研究分野における未解決問題の一つである異なる二つの系統のウイルスが互いに興隆と没落を繰り返しながら流行するような 2 系統のウイルス流行の同時予測問題を考察し, 相互作用をとり入れた数理モデルを提案している. この 2 系統予測問題は, インフルエンザワクチンの設計で必要な翌年の流行に対する優勢株の予測に応用をもち, 感染症分野において重要な問題の一つである. 二つ目の柱は, データ同化の枠組みにおける数理モデルのパラメータ推定である. 一般に複雑な時系列モデルでは, システムのパラメータに関する不安定性や隠れ変数の存在によって, 対象分野特有の事前知識だけではパラメータ設定が難しい. この問題に対して著者らは, 与えられた観測データから統計推論を用いてパラメータ値を推定するデータ同化の枠組みを採用し, 近似ベイズ推論 (ABC, approximate Bayesian computation) に基づくパラメータの推論手続きを与えている. 第三の柱として, データ同化システムの高速化について考察し, 実際に流行予測システムを大規模観測データに適用する際のボトルネックとなる空間データアクセスの高速化に関して, 空間索引の一種である分散 4 分木 (distributed quadtree) を用いた二つの高速化手法を提案している. さらに, 上記の提案の有効性を実際のデータを用いた計算機実験により検証している.

具体的には, 本論文の各章では以下の問題について研究している.

3 章では, 異なる 2 系統のインフルエンザ B ウイルスの流行の同時予測問題を考察し, そのための数理モデルと推論手続きについて研究している. 1 系統の時系列予測モデルとして, S, E, I, R の 4 つの状態変数をもつ微分方程式系である SEIR モデルが広く用いられている. この 2 系統予測への拡張として, 二つの独立した SEIR モデルを用いる素朴なモデルが考えられるが, 著者による予備実験ではこの素朴なモデルは低い予測精度しかもたないことが判明した. そこで著者は, 異なる系統間での相互作用の項を追加して, 2 系統予測のための「積 SEIR モデル」 (the product SEIR model) を提案している. 提案モデルは尤度関数を明示的に求められない数理モデルであるが, 著者は 近似ベイズ推論に基づくパラメータ推論手続きを与えることでこの問題を解決し, 2 系統のインフルエ

ンザ B ウイルスの流行予測システムを実現している. さらに, 著者は Yamagata 種と Victora 種の 2 系統のインフルエンザ B ウイルスの流行に関する実データを用いて評価実験を行い, 提案手法が, 翌年に優勢となる系統の予測に必要な, インフルエンザ B ウイルスの主要なパラメータを推測することに成功していると結論している.

4 章と 5 章では, データ同化を用いた予測システムにおけるデータ同化におけるデータアクセスのボトルネックの軽減方法に関して考察している. さらに評価実験を行い, データ同化において必要な空間データアクセスに関して, 提案する手法が大幅な高速化に貢献することを実証している. はじめに, 4 章では, メモリ非共有型の分散計算機システムにおいて, 分散 4 分木の利用による空間データアクセスの高速化手法について考察し, 分散ノードへのデータとクエリの分割と割付について, 分散計算機上の実装を用いて各種の手法を比較している. 実験結果からは, 単一の大きな外部索引を用いる従来のデータアクセス方式に対して, 分散計算機上の 4 分木を用いた提案方式の優位性が観察されている. 次に, 5 章では, さらなる高速化のために, 分散索引におけるデータの格納時と検索時のクエリ転送計算の高速化を考察している. 4 分木構築に静的サンプルを用いることを利用して, ハッシュ表 (hash map) と座標の整数符号化 (dilated integer coding) を用いて, 転送データ索引の検索を高速化する手法を提案している. 実験では, 提案手法が従来手法に対して約 3 倍高速であり, データの偏りに対してより安定であることが示されている.

最後に 6 章では, 本論文の結果と貢献をまとめ, 本研究の今後の方向性を与えている.

これを要するに, 著者は, 時系列データ解析における感染症の流行予測の領域において, 数理モデル設計と, 推論手続き設計, 高速化および, 実データを用いた実証的評価に基づいて, 大規模観測データと数理モデルを用いた高精度で効率良い時系列予測が可能であるという新知見を得たものであり, 情報科学における大規模データ処理分野と生物情報学分野に貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.