



HOKKAIDO UNIVERSITY

Title	The Analysis of Infectious Diseases via Machine Learning [an abstract of dissertation and a summary of dissertation review]
Author(s)	TESSMER, HEIDI LYNN
Degree Grantor	北海道大学
Degree Name	博士(獣医学)
Dissertation Number	甲第13258号
Issue Date	2018-06-29
Doc URL	https://hdl.handle.net/2115/71244
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/4.0/
Type	doctoral thesis
File Information	Heidi_Lynn_TESSMER_abstract.pdf, 論文内容の要旨



学位論文内容の要旨
Abstract of the dissertation

博士の専攻分野の名称：博士（獣医学）

氏名：Heidi Lynn Tessmer

Name

学位論文題名

The title of the doctoral dissertation

The Analysis of Infectious Diseases via Machine Learning

(機械学習による感染症の解析)

This thesis introduces two projects applying machine learning methods to the realm of bioinformatics. In Chapter 1, we look at a regression problem involving the parameter values associated with the SEIR epidemiological model while in Chapter 2 we explore viral host classification.

Chapter 1 - To estimate and predict the transmission dynamics of respiratory viruses, the estimation of the basic reproduction number, R_0 , is essential. Recently, approximate Bayesian computation methods have been used as likelihood free methods to estimate epidemiological model parameters, particularly R_0 . In this paper, we explore various machine learning approaches, the multi-layer perceptron, convolutional neural network, and long-short term memory, to learn and estimate the parameters. Further, we compare the accuracy of the estimates and time requirements for machine learning and the approximate Bayesian computation methods on both simulated and real-world epidemiological data from outbreaks of influenza A(H1N1)pdm09, mumps, and measles. We find that the machine learning approaches can be verified and tested faster than the approximate Bayesian computation method, but that the approximate Bayesian computation method is more robust across different datasets.

Chapter 2 - Infectious diseases which transfer between species are particularly difficult to manage. Knowing the natural host for an infectious agent makes it easier to prevent interspecies transmissions. However, with new and re-emerging disease, it can be difficult to know what the reservoir host is. In the second half of this thesis, we conducted a principal component analysis using data from the fruit bat and wild duck, along with a selection of single-stranded RNA viruses found in each animal. Historically, the virus-host relationship has often been examined using two components, that is, the G+C content of the genomes and the rate ratio of CpG in the genome. However, numerous data discrepancies exist which cannot be explained with mathematical models built from this technique. In this study, we found several alternative components that could be used to infer the host animal species of RNA viruses. Using these alternative components, we may be able to build a mathematical model that more closely simulates the virus-host genetic relationship. With this information, we may be able to identify genetic signatures in viruses which can uniquely identify the natural host species. In future, this information could help identify the animal source of a new outbreak.