大规模函数型数据分析存储空间与计算效率问题取得进展

作者：writer 来源：科学网

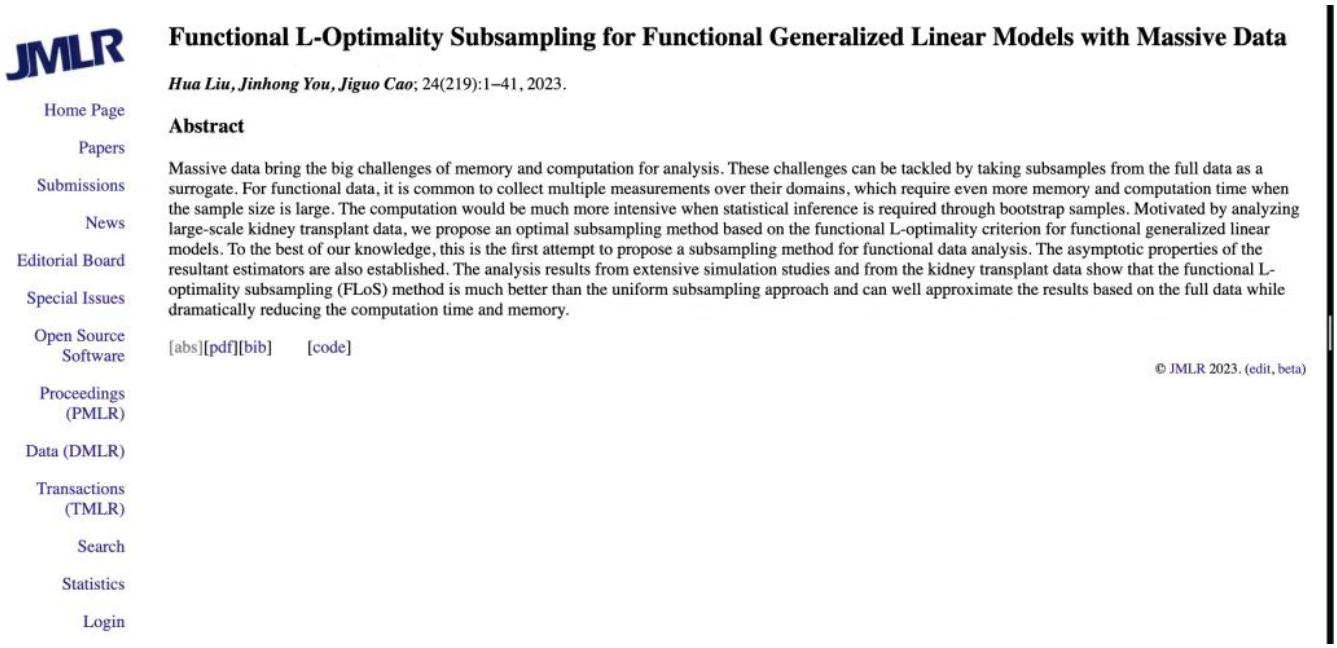
本文原地址：https://www.iikx.com/news/progress/23969.html

大规模函数型数据分析存储空间与计算效率问题取得进展。

在大数据时代下，随着互联网、云存储等技术的高速发展，实际分析处理中遇到的数据规模越来越大。尽管大规模函数型数据可以给我们带来海量信息，但是其对计算资源有着更高的需求，所需的计算时间更长，这也极大地提升了计算成本，影响数据分析的时效性、可操作性。因此如何解决大规模函数型数据分析时，遇到的存储空间和计算效率等方面的难题是大数据时代下函数型数据分析的一个重要问题。

近日，针对上述问题，西安交通大学经济与金融学院青年教师刘华博士、上海财经大学统计与管理学院教授尤进红博士和加拿大西蒙弗雷泽大学教授Jiguo Cao博士进行了深入的研究。他们首次把子抽样思想应用到函数型数据分析，开发出适应于函数型广义回归模型的最优抽样方法Functional L-Optimality Subsampling(FLoS)，以此来实现减少计算时间、克服内存不足等问题的目标。除此之外，作者还通过理论和一系列的数值模拟来说明了该抽样方法的准确性和有效性。

科研人员把提出的最优抽样方法FLoS用于分析器官移植数据案例，该数据收集了几十万名肾脏器官移植接受者在接受器官移植手术时的信息，并记录了这些移植手术接受者在术后每次随访的信息，因此其是一个数据量非常庞大的且包含函数型数据的数据集。他们想要用接受者术后的肾小球过滤率曲线来判断移植手术能否成功并且预估他们在术后的大致寿命。通过分析与对比，他们发现基于FLoS方法抽取到的最优子样本得到的抽样估计和全样本下的估计几乎完全一致，进一步验证了该最优抽样方法的准确性和有效性。



JMLR

Home Page
Papers
Submissions
News
Editorial Board
Special Issues
Open Source Software
Proceedings (PMLR)
Data (DMLR)
Transactions (TMLR)
Search
Statistics
Login

Functional L-Optimality Subsampling for Functional Generalized Linear Models with Massive Data

Hua Liu, Jinhong You, Jiguo Cao; 24(219):1–41, 2023.

Abstract

Massive data bring the big challenges of memory and computation for analysis. These challenges can be tackled by taking subsamples from the full data as a surrogate. For functional data, it is common to collect multiple measurements over their domains, which require even more memory and computation time when the sample size is large. The computation would be much more intensive when statistical inference is required through bootstrap samples. Motivated by analyzing large-scale kidney transplant data, we propose an optimal subsampling method based on the functional L-optimality criterion for functional generalized linear models. To the best of our knowledge, this is the first attempt to propose a subsampling method for functional data analysis. The asymptotic properties of the resultant estimators are also established. The analysis results from extensive simulation studies and from the kidney transplant data show that the functional L-optimality subsampling (FLoS) method is much better than the uniform subsampling approach and can well approximate the results based on the full data while dramatically reducing the computation time and memory.

[abs][pdf][bib] [code]

© JMLR 2023. (edit, beta)

研究成果发表在JMLR.(图源JMLR网站)

近日，上述研究成果以《大规模函数型广义回归模型下的最优抽样方法FLoS》为题发表在机器学习和人工智能领域国际顶级期刊Journal of Machine Learning Research(简称JMLR)上。刘华是第一作者，西安交通大学经济与金融学院是第一署名单位。JMLR由麻省理工学院出版社出版，依托于麻省理工学院的计算机科学与人工智能实验室，是国际上公认的计算机领域顶级期刊之一。(来源：中国科学报 严涛)

相关论文信息：https://www.jmlr.org/papers/v24/22-0614.html

作者：刘华等 来源：《机器学习研究杂志》