JOHNS HOPKINS
SHERIDAN LIBRARIES
Open Source Programs Office

Project Overview

Project Title

Driving precision medicine forward: large-scale transfer learning for multi-institutional health data via extension of the BayesBridge package (github.com/ohdsi/bayes-bridge)

Project Summary

Patients' interactions with healthcare systems are increasingly captured and made available for research purposes in the form of electronic health records (EHRs) and insurance claims. The hope is to improve healthcare practices by turning these data into actionable insights. Many clinical questions of interest are beyond the reach of any single data source, however; some data sources may cover a large number of patients but lack breadth and depth of information, while the opposite holds for others. This has given rise to consortiums of distributed health databases, realizing the full potential of which require a method to synthesize information across databases. Bayesian modeling excels at such tasks, but the scale and distributed nature of these health data creates an obstacle to deploying existing Bayesian software. We extended the state-of-the-art BayesBridge package and implement a skew-shrinkage regression model for scalable transfer learning across distributed databases.

Target Audience

The BayesBridge package forms a part of the HADES software suite (ohdsi.github.io/Hades) and its earlier version has been widely used by researchers in the Observational Health Data Sciences and Informatics (OHDSI) consortium. The new skew shrinkage feature addresses one of OHDSI' most urgent needs: to transfer information obtained from training a model on a larger database to a smaller database, without sharing patient level data. As such, the new feature is expected to serve as a critical component in realizing the full potential of OHDSI's distributed health data network.

Code Repository

The code is available in the `skew-shrinkage` branch of the GitHub repository for the BayesBridge package at https://github.com/OHDSI/bayes-bridge/tree/skew-shrinkage.

Project Activities and Progress

Work Completed

We have successfully completed implementing the proposed skew-shrinkage model for large-scale transfer learning as a new feature within the BayesBridge package. We are in the process of demonstrating gains in predictive ability using datasets from IQVIA Pharmetrics Plus (source data) and Johns Hopkins EHR (destination data).

Technical Milestones

We started the implementation of the new feature by refactoring the code to make it easier to use alternate priors (and, in particular, the proposed skew-shrinkage prior) for Bayesian sparse regression. We then implemented, and tested them as independent modules, the two new sampling algorithms necessary to carry out posterior inference under the skew-shrinkage prior. Getting these sampling algorithms right took several iterations and debugging for the student trainees. These sampling algorithms were then refactored and integrated into the main posterior inference algorithm.

Challenges and Solutions

When the earlier versions of code yielded unexpectedly poor results on real datasets, the inherently stochastic/statistical nature of the implemented method made it difficult to attributes the poor performance to the method itself or bugs in the code. We were able to identify the bugs through simulation studies using synthetic data, where the method behavior was more a priori predictable.

Outcomes and Impact

Project Impact

The large-scale transfer learning tool is scheduled, after we complete the initial validation and demonstration on the real-world datasets in next few months, to be validated on a broad range of observational health databases across the extensive set of benchmark metrics. This validation is part of the standard practice OHDSI has established to ensure scientific validity of clinical evidence generated by our data science pipelines and all newly proposed methods must go through before they can be deployed in clinical applications. While its practical values are yet to be established, the tool has nonetheless been highly anticipated by the Method Research team and the broader OHDSI community.

The tool also has a potential application to the problem in statistical genetics of predicting disease risks based on individuals' genotype. It can be applied to transfer some of the information from the model trained on European populations, where a large amount of genotype data is available, to help train the model for African populations for whom a relatively small amount of data is available. We plan to integrate the tool into *PRS-Bridge* (https://github.com/YuzhengDun1999/PRSBridge), a fork of BayesBridge adapted for the genetic disease risk prediction problem.

Community Engagement

The new skew-shrinkage feature for large-scale transfer learning will be presented at upcoming 2024 OHDSI Global Symposium in New Brunswick, NJ. The last symposium garnered over 440 attendees and we expect comparable or greater level of attendance this year. The feature is also scheduled to be presented at a future meeting of the OHDSI's Method Research Working group.

Sustainability / Future Plans

I used statistical results based on the work completed under the FOSSProf support as preliminary data in my application for NIGMS's Maximizing Investigators' Research Award. The software package will be maintained and further developed to support more features for both research purposes and clinical applications. The two student trainees of this project will continue contributing to the maintenance and development of the package.

Lessons Learned

The implemented method has demonstrated significant improvements in prediction for some applications but relatively smaller improvements in others. While there is little doubt that the tool provides value overall, the limited improvements seen for some of the settings made me realize a challenge for software development somewhat unique to the academic research setting. We often propose novel methods that have never been implemented before; as such, these methods are not of proven values no matter how promising they are. Such uncertainty in value of new features is also true of software development more generally, but the nature of academic research does mean we are often pursuing more experimental and high-risk propositions. This makes rapid prototyping and evaluation important. On the other hand, a thorough evaluation warrants reliable implementation. Careful balancing of development speed and code quality/reliability, while accounting for the unique requirements of academic research, is thus essential.

Attachments

Code snippet, as an HTML rendering of a Jupyter notebook, to demonstrate the use of the new skew-shrinkage feature is attached.