Doctor of Philosophy in Bioinformatics
Abstract of the Doctoral Thesis

Development of Privacy-preserving Machine Learning Methods for Metabolomics
（メタボロミクスにおける プライバシー保護機械学習法の開発）

Graduate School of Science and Technology
Degree Programs in Systems and Information Engineering
Doctoral Program in Life Science Innovation: Bioinformatics

Akihiro MIZOGUCHI

March 2023

Abstract

Metabolomics seeks to elucidate biological phenomena via a comprehensive analysis of metabolites. It is considered to be a more downstream area of study, compared to genomics, which analyzes genes; transcriptomics, which analyzes gene expression; and proteomics, which analyzes proteins. This is because metabolites are most closely associated with phenotypes. Thus, metabolomics has been used to diagnose diseases, discover new biomarkers, and identify new drug candidate compounds.

Because metabolomics involves a large number of metabolites, reliance on the use of machine learning methods to analyze relevant objectives is gaining momentum. For example, machine learning models that are able to predict disease phenotypes based on the concentrations of certain metabolites in blood, or toxicity based on the known structure of a particular metabolite, have been constructed. Generally, the greater the number of data used for training, the higher the predictive accuracy of machine learning. With respect to metabolomics, compound data pertaining to metabolite concentrations gathered from multiple institutions engaged in machine learning have helped improve predictive accuracy much more than data gathered only from a single institution engaged in machine learning.

Privacy-preserving machine learning methods, such as federated learning and data collaboration analysis, were developed to resolve the above stated issues. However, several difficulties are encountered when attempting to apply these privacy-preserving machine learning methods to metabolomics. These include the misalignment of features among multiple institutions as well as the low performance of distributed data, which exhibit a high degree of label bias. Federated learning, which is highly versatile, has been applied in a wide range of fields, such as mobile devices, diagnoses based on medical images, and prediction of the properties of compounds developed by pharmaceutical companies. However, it has not been able to overcome certain challenges.

In this dissertation, we describe our attempts to address these challenges using techniques such as the extension of data collaboration analysis. Data collaboration analysis was proposed as a privacy-preserving machine learning method by Imakura and Sakurai (2019). This method can be used to transform raw data distributed across multiple organizations into an intermediate representation via a dimensionality reduction of each organization and store it on a server for machine learning, which enables machine learning using all data without sharing raw data. Data collaboration analysis, which has been applied to tabular data as well as image data, was compared with federated learning, which has been applied to image data, in an independent and identically distributed (IID) setting. However, to the best of our knowledge, no studies have applied data collaboration analysis to distributed data with misaligned features, or to non-IID settings. Therefore, we aimed to address the challenges of limited data partitioning and low performance in non-IID settings by extending data collaboration analysis and applying it to distributed data with misaligned features or those in non-IID settings.

First, we addressed the issues pertaining to limited partitioning of metabolite concentration data.

Metabolite concentration data distributed across multiple studies and institutions are expected to have different characteristics. However, federated learning and previous data collaboration analysis have not been applied to data partitioning, which involves completely different samples with only some common features, a situation that is often encountered when partitioning metabolite concentration data.

Based on this background, we extended data collaboration analysis in a manner that enabled it to be applied to data partitioning of completely different samples with some common features, which had not been previously subjected to federated learning or data collaboration analysis. Specifically, we devised a method for creating anchor data that allowed distributed data with misaligned features to be converted into a common intermediate or collaboration representation. To perform this evaluation, we created four artificial datasets and two datasets generated from public metabolomic data containing completely different samples with partially common features. The machine learning performance resulting from data collaboration analysis using common features as well as non-common features was higher than that resulting from conventional data collaboration analysis using only common features or individual analyses.

However, we only showed that the proposed method was useful for artificially generated data containing completely different samples with partly common features. We plan to apply the proposed method to real-world metabolomic data derived from different measurement methods, hospital patient data, and bank customer data.

Next, we investigated the low performance observed in the non-IID settings of compound datasets, where the distribution of labels in each user's data is highly skewed. With respect to predicting the properties of a compound based on its structure, the distribution of labels in the data tends to be skewed as a whole, due to the number of compounds exhibiting efficacy or toxicity being smaller than of those without these properties. In addition, it is possible that the distribution of labels in pharmaceutical company data is skewed, because each company has highly variable compound data. Previous research indicates that federated learning shows low machine learning performance in non-IID settings.

Therefore, we examined the robustness of data collaboration analysis in non-IID settings and proposed a new method to improve previous data collaboration analyses in non-IID settings, via the application of data collaboration analysis to compound data. First, we applied data collaboration analysis to compound data obtained from the PubChem compound database as anchor data, instead of applying it to random values as previous data collaboration analyses have done. Then, we introduced the concept of projection data, which creates intermediate representations for data collaboration analysis, thereby improving machine learning performance in non-IID settings and proposed a method termed data collaboration analysis with projection data (DCPd). We then compared classification performance using area under the receiver operating characteristic curve (ROC-AUC) and area under the precision-recall curve (PR-AUC) of federated averaging (FedAvg), data collaboration analysis (DC), and DCPd, in both IID and non-IID settings, using five compounds datasets, CYP2D6, CYP3A4, CYP1A2, HIV, and

Tox21_SR-ARE, derived from Therapeutic Data Commons (TDC). The results showed that the classification performance of the three methods was very similar in IID settings, whereas DCPd showed a higher classification accuracy compared to FedAvg and DC in non-IID settings. A subsequent experiment showed that the classification performance of DCPd did not decrease significantly with increasing label bias of each user's data whereas that of DC decreased slightly and that of FedAvg dropped rapidly.

The results of this study enabled data collaboration analysis to be applied to the prediction of compound properties and confirmed the robustness of data collaboration analysis in non-IID settings. It also showed that the proposed method, DCPd, enhances this robustness. Moreover, it may be desirable to make data collaboration analysis applicable to graph data, because graph convolutional networks and other methods that use graph structures as input for property prediction of compound data have become mainstream in recent years. Furthermore, the proposed method should be compared with improved methods of federated learning in non-IID settings in the future.

As stated before, this dissertation demonstrates that our extended data collaboration analysis method would be capable of addressing two of the challenges that are encountered when using privacy-preserving machine learning in metabolomics. These challenges are (i) limited data partitioning and (ii) low performance in non-IID settings. However, other challenges that need to be addressed, such as privacy concerns, remain. Additionally, to broaden the application potential of data collaboration analysis, certain challenges specific to this method, such as inability to handle graphic structures and natural language data, and the method of creating anchor data, must be addressed.