A novel privacy-preserving software framework for federated machine learning in clinical settings

Marta Lemanczyk, Dominik Heider

Department of Bioinformatics, Faculty of Mathematics and Computer Science, Philipps-Universität Marburg, Germany

Sharing of patient data between cooperation partners is an important component of biomedical research as prediction models can be improved by using a larger amount of data. Due to privacy regulations, such as the General Data Protection Regulation (GDPR) by the EU (2018), sharing of sensitive biomedical data of patients is not possible without permissions. Thus, keeping data centralized and making it only available for cooperation partners is not an option since it would increase the risk of cyber-attacks. Furthermore, anonymization of individual data is in most cases not sufficient. It has been shown that it is possible to trace back the identity of one patient by reproducing it from data [1]. However, collaborations between institutions require a secure solution for sharing information gained from patient data.

Federated Machine Learning (FML) enables information exchange between users by sharing parameters and meta-information of machine learning models, without violating the privacy requirements. Each user stores their own data locally and has no access to other users data. Without sharing any sensitive information, a user can transfer parameters and meta-information of a locally trained model and can combine it with the other models towards a more generalized model.

In this study, we focused on the development of a secure privacy-preserving software framework, which can be used for developing federated machine learning models. Moreover, we analyzed multi-party model combinations. Depending on data size and class balance, different combinations of models were evaluated, with a particular focus on logistic regression and random forests [2], due to the fact that these models are well-accepted in the biomedical community. For training and evaluation of the models, two real-world data sets were used, namely diabetes prediction [3] and prediction of antimicrobial efficacy [4]. Additionally, the concept of differential privacy [5] is applied on the models to prevent reproduction of patient identities.

References

[1] Sweeney L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002, 10(5): 557-570.

[2] Breiman L.: Random forests. Machine learning 2001, 45(1): 5-32.

[3] Kälsch J., et al.: Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort. Scientific reports 2015, 5: 13058.

[4] Pirtskhalava M., et al.: DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. Nucleic acids research 2015, 44(D1): D1104-D1112.

[5] Pathak M., Shantanu R., and Bhiksha R.: Multiparty differential privacy via aggregation of locally trained classifiers. Advances in Neural Information Processing Systems 2010, 23:1876-1884.