The Batch Effect and the Necessity of Metadata Standardization

Scott Dierckman
Major: Actuarial Science

Eric Frempong
Major: Mathematical Sciences

Nils Hinniger
Major: Actuarial Science

Stephen Owuso
Major: Mathematical Sciences

Brandon Rigdon
Major: Actuarial Science

Morgan Scalpone
Major: Actuarial Science

William Sellers
Major: Mathematical Sciences

James White
Major: Actuarial Science

Alexandre Yano
Major: Mathematical Sciences

Faculty Sponsor and co-author: Dr. Alessandro Selvitella
Department: Mathematical Sciences

Faculty Sponsor: Dr. Todor Cooklev
Department: Electrical and Computer Engineering

The Batch Effect (BE) is the phenomenon in which noisy variables such as the type of equipment and/or the type of metadata standard, confound the signal observed. The BE is relevant to Radio Frequency (RF) spectrum analysis, as RF samples are taken and processed with different protocols, different instrumentations, and possibly different metadata representations. Metadata have proven to be crucial in the analysis of the RF spectrum, but because the spectrum monitoring is performed using platforms with different capabilities and parameters, it opens the door to the BE. Integrating platforms with different metadata standards and converting between metadata representations is a fundamental problem, which is still largely open. We propose the implementation of ML methods to address the problem of BE for RF metadata. A typical situation is the following. If the metadata with Standard A is different from the metadata with Standard B, ML methods can learn a conversion standard, namely a map (in general highly nonlinear, like a deep neural network) that gives a representation of the metadata of Standard A as seen by Standard B (or vice-versa). In this way, ML can become a valuable part of the interface. Note that in the absence of standardization, particular care needs to be taken when combining data accompanied by different metadata standards. Aggregated data can potentially facilitate statistical analyses, but there is a catch: the metadata of different samples need to be compatible, for the aggregated data to be useful. In the absence of standardization, this cannot be assumed, and so any statistical assessment would suffer from higher uncertainty. A preliminary evaluation of compatibility between metadata standards can be made using ML. Suppose we try to solve a classification problem where the data is the predictor (input) and

the metadata standard is the outcome variable (output). Having an algorithm with a high classification accuracy in this task would mean that the signal is confounded by the variability of the metadata standard. This situation would call for metadata standardization, because it implies that the signal can be distinguished by means of the metadata standard used, which translates to a lack of interoperability between systems. The detection of a strong BE would support the need of metadata standardization. ML can quantify the BE, assess dataset compatibility, and provide a conversion between metadata standards. The adoption of a common metadata standard will facilitate interoperability and prevent technology segmentation.

In this work, we are analyzing simulated data of RF spectrum monitoring. Issues associated with the compatibility of the data format, transmission and processing of the data, and determination of the most suitable software for data processing are analyzed.

This project is directly connected to Prof. Selvitella's *Spring 2020 START Initiative Grant* on "*Deep Learning with Applications to Wireless Communication*" and it is of specific interest to the Center for Applied Mathematics and Statistics, Purdue University Fort Wayne, and Stryke Industries.