DOI:10.3217/978-3-85125-260-6-180

Bringing Big Data to Neural Interfaces

I. Obeid[1], J. Picone[1]

[1]*Temple University, Philadelphia, Pennsylvania, USA*

Correspondence: I. Obeid, Temple University, Department of Elect. and Comp. Engineering, 1947 N. 12th Street, Philadelphia PA 19122, USA.
E-mail: iobeid@temple.edu

Abstract. The purpose of this paper is to present a new community-wide research entity to be launched called the Neural Engineering Data Consortium (NEDC). The purpose of the NEDC is to accelerate Brain Computer Interface research by creating, curating, and archiving massive neural datasets for the neural engineering community. The need for massive datasets is justified by the innate variability in neuronal activity. By pooling the resources of the neural engineering community and making such massive datasets available, we will enable investigators to improve BCI performance metrics and increase robustness. A proof of concept data corpus comprising 12,000 clinical EEGs is presently under development.

Keywords: Big Data, Machine Learning, EEG, Data Mining

1. Introduction

The past two decades have seen an explosion in Brain Computer Interface research dedicated to using neural signals for controlling prosthetic interfaces to the external world. However, despite significant progress in recent years, overall progress in the field does not appear to have been commensurate with the scope of investment (over $200M in the last decade from NIH and NSF alone). In particular, efforts to commercialize research findings have been tepid, hampered by a general lack of robustness when translating technologies to uncontrolled environments beyond the research laboratory. It has become evident that the field would benefit from a new paradigm in research and development that focuses on robust algorithm development.

In response to the current state of Neural Engineering research, we are launching the *Neural Engineering Data Consortium* (NEDC). The purpose of this organization is to focus the attention of the research community on a progression of neural engineering research questions and to generate and curate massive data sets to be used in addressing those questions. The existence of massive common data corpora has proven to substantially accelerate research progress by eliminating unsubstantiated research claims. The NEDC will also broaden participation in neural engineering research by making data available to research groups who have significant signal processing expertise but who lack capacity for data generation. This effort is modeled in part after similar successful endeavors, particularly in the human language technology field where a data consortium has led to systematic research and technology advances over a 20-year span.

2. NEDC Structure and Role

2.1 Overview

The NEDC is designed as an independently operated community resource with the goal of providing some measure of value to research labs running the gamut from the largest and most well established groups down to small unfunded labs. The NEDC does not seek to compete for funds with other members of the community, nor does it seek to supplant the existing roles of any members of the community or undermine their autonomy. Rather, we envision the NEDC as a communal utility whose actions will serve the greater good. The NEDC operates independently under the auspices of Temple University.

Working with a group of community stakeholders, the NEDC will be developing a series of research questions that can be addressed with a series of progressively difficult tasks. Various federal funding agencies will then contract the NEDC to design and execute data collection protocols. A critical role of the NEDC is to design data collection protocols that account for all the various biases that can compromise complex data models and that are large enough to accurately account for data variances. In the case of neural data, this may well require corpora that are orders of magnitude larger than what any single PI or group could realistically collect.

Data generated by the NEDC will be released to the consortium members, including both funded and unfunded PIs, as well as members of industry. Investigators will pay a tiered consortium fee to secure access to the data. The value of giving unfunded investigators discounted access to data has been well documented; many of the best contestants in the Berlin Brain-Computer Interface contests were small laboratories without the means to generate their own brain interface data [Blankertz et al., 2006]. Providing labs such as these with access to data is therefore a means of significantly improving the community's overall progress at little or no effective cost.

2.2 Data-Driven Research Challenges

Perhaps the most critical aspect of the NEDC's mandate is to issue research challenges and to independently score and arbitrate the results in much the same way as the Berlin Brain-Computer Interface contest has. Contestants would use a set of training data to create a model, and would then use that model to decode a set of test data. The NEDC would then score the results and periodically announce ranked winners. This is a vital aspect of the NEDC because it makes it possible for research labs to compare performance on the very same data. In this sense, it will be possible to disambiguate the best modeling and neural decoding methods from less promising approaches. This is not possible under the existing research model, in which each lab uses its own data to test its algorithms; even existing shared data sets are not sufficiently rich to adequately evaluate any given neural decoding algorithm.

3. EEG Data Corpus

The NEDC is presently creating its first data corpus as a proof-of-concept to demonstrate the value of big data as well as our own operational capabilities. Using so-called "found data," we are creating what will be the world's largest publicly available database of clinical EEG data. The data comprise over 12,000 clinical EEG records made at Temple University Hospital over a ten-year period. Although information disclosing a patient's identity, such as name and corresponding video will be redacted, other demographic information such as gender, age, ethnicity, relevant medical history, and medications will be retained. In this manner, for example, it will be possible to mine the completed data set for statistically significant changes in EEG activity in response to various medications. In this regard, the NEDC is not only creating a tool of relevance to biomedical community, but is also developing and demonstrating our capability for managing and disseminating data of the magnitude we envision for the future. The EEG dataset is expected to be complete and available by the end of 2013.

4. Discussion

Many of the engineering challenges facing the BCI community such as high variance signals and incremental progress in improving signal processing performance are common to other engineering fields; lessons learned in those areas suggest ways forward for BCI investigators. In the human language technology (HLT) area, these engineering challenges have been effectively managed by the creation of the Linguistics Data Consortium (LDC). The LDC (*www.ldc.upenn.edu*) was founded 20 years ago in response to a realization that progress in speech processing was being impeded by a lack of common data sets and a lack of common research goals. The lack of common research goals fragmented the community and made it difficult to make meaningful inroads into any one research topic. The LDC was launched with support from NSF and DARPA and, although an independent entity, is managed administratively through the University of Pennsylvania.

The LDC performs three principal functions: (1) defining research problems of interest to the community at large (2) designing and executing data collection protocols to create massive data corpora and (3) distributing the data to consortium members and then, in collaboration with NIST, independently verifying signal processing performance claims from the competing laboratories. HLT research demands large data sets that are often beyond the capacity of any one organization to collect. These data sets are critical to the sophisticated machine learning algorithms employed in state of the art systems. As a result of LDC's influence, HLT performance has been steadily improving as datasets have increased in size from minutes of data to tens of thousands of hours of data. Further, these vast data sets have enabled the development of sophisticated unsupervised learning paradigms that have dramatically decreased the cost of developing systems.

The BCI field is ready for this type of innovation. The NEDC hopes to lead the way by creating such synergies within the bioengineering community. NEDC will collaborate closely with LDC so that we can build on their best practices and leverage their considerable expertise in experimental design, data annotation and distribution.

References

Blankertz B, Müller KR, Krusienski DJ, Schalk G, Wolpaw JR, Schlögl A, Pfurtscheller G, Millán JdR, Schröder M, Birbaumer N. The BCI competition III: Validating alternative approaches to actual BCI problems, *IEEE Trans Neural Syst Rehabil Eng*, 14(2), 2006.