Truthiness: Challenges Associated with Employing Machine Learning on Neurophysiological Sensor Data

Mark Costa[1,2](✉) and Sarah Bratt[2]

[1] School of Information Studies, Syracuse University, Syracuse, USA
[2] M.I.N.D. Lab S.I. Newhouse School of Public Communications, Syracuse University, Syracuse, USA
{mrcosta,sebratt}@syr.edu

Abstract. The use of neurophysiological sensors in HCI research is increasing in use and sophistication, largely because such sensors offer the potential benefit of providing "ground truth" in studies, and also because they are expected to underpin future adaptive systems. Sensors have shown significant promise in the efforts to develop measurements to help determine users' mental and emotional states in real-time, allowing the system to use that information to adjust user experience.

Most of the sensors used generate a substantial amount of data, a high dimensionality and volume of data that requires analysis using powerful machine learning algorithms. However, in the process of developing machine learning algorithms to make sense of the data and subject's mental or emotional state under experimental conditions, researchers often rely on existing and imperfect measures to provide the "ground truth" needed to train the algorithms.

In this paper, we highlight the different ways in which researchers try to establish ground truth and the strengths and limitations of those approaches. The paper concludes with several suggestions and specific areas that require more discussion.

Keywords: Machine learning · Cognitive data · Method validity · fNIRS · Neurophysiological sensors

1 Introduction

The goal of using machine learning on data generated by neural sensors is to predict or identify a user's mental state in real-time. The roadmap to achieving that goal usually involves conducting controlled experiments, where subjects are exposed to a series of control and treatment conditions. Depending on the experimental setup, the subject's mental state under the treatment condition can be either identified through self-report measures, the nature of the task and its expected effect on mental state, or the subject's task performance. The labels are then used as predictor variables for machine learning

Costa, et al. p. 1, 2016

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part I, LNAI 9743, pp. 159–164, 2016.
DOI: 10.1007/978-3-319-39955-3_15

algorithms. Each of these approaches has drawbacks in terms of validity and reliability, which may lead us to train the algorithms on incorrectly labeled data (Hoskin 2012).

In an ideal world, the problems of mislabeled data would average out over large data sets – machine learning is successfully applied in many spaces where a high volume of data is available to train the models. However, in the applied neuroscience space the number of cases is usually very small and the dimensionality of the data is very large, which can easily lead to overfitted models. This is one of the reasons why developing models that work across subjects, experimental conditions, and/or treatments is very difficult.

To sum up the challenge in a single sentence, we are trying to build predictive models on unreliably tagged data under the curse of dimensionality.

There are valid reasons for undertaking this effort. First, these efforts will enable future systems that adapt to our mental, physical, and emotional state in real-time, helping us make better decisions, gain deeper insights, and solve bigger problems, from medical diagnoses to adaptive military technology (Gateau et al. 2015; Girouard 2010; Naseer and Hong 2015; Marx et al. 2015). Developing such systems will involve integrating voluminous data from multiple sensors, a task which machine learning is especially well-suited. This paper addresses one of the handful of challenges associated with building adaptive systems – identifying the ground truth to build upon when developing models and systems.

The remainder of the paper is structured as follows: first, a brief description of neurophysiological sensors is provided. Then, an overview of the approaches to labelling data for algorithm training purposes, and a discussion of the validity and reliability problems associated with the labeling approaches follows. The paper concludes with a discussion of potential solutions and directions for future research.

2 Generating Data

Neurophysiological sensors rely on different physical mechanisms to measure activity in the brain. For example, an EEG measures electrical activity generated by neurons firing within the brain, fNIRS measures blood flow to and from areas of the brain instigated by activation and deactivation of specific regions, and fMRI uses magnetic resonance imaging to track blood flow in a manner similar to fNIRS.

The main point to consider when thinking about labeling neurophysiological data for machine learning purposes is that the sensor generates one row of data every time it samples. For example, an fNIRS can be set to sample at 10 Hz, generating approximately 36,000 rows of data for a one hour experiment. For each row there may be some number of data points associated with the channels in the device, which we can call features. Supervised learning algorithms use those features and derivative features to build a model that predicts a label. Most algorithms perform this model creation and evaluation by passing over the data frequently, iteratively refining the weights given to each of the features until such a point that the algorithm has satisfied its optimization criteria. The labeling process involves estimating the subject's mental or emotional state at each of those points in time and assigning a category code (the label, or “class label”) to that point in time or interval of time.

The labeling process results in at least a few error in the labeling process in the boundary regions between state shifts because identifying the exact point in time when a cognitive state changes in not currently possible. Furthermore, we argue in this paper that the labels in general are not entirely reliable due to limitations of the approaches available, and as a result, the trained models are not reliable. Using incorrectly labeled data to train a supervised algorithm would be analogous to teaching a child to add by giving her a set of addition problem examples with answers that were correct only some of the time, then expecting the child to know how to add when given a new set of problems.

Before going into detail justifying our argument, we start with a brief description of the approaches researchers have used to determine those labels and the justifications for making that choice (Fairclough and Gilleade 2014; Noah et al. 2015; Liu et al. 2015).

3 Approaches to Labeling Data

We identified three approaches to labeling the data – response based, task based, and task-performance based. Here we will refer to the label as “ground truth” – what the researcher believes to be the best approximation of the subject’s mental or emotional state. Each approach has strengths and weaknesses, and none appears to be the superior approach.

3.1 Response-Based Labeling

Response based labeling uses the subject’s subjective interpretation of their mental state as the ground truth. For example, researchers have used the self-assessment manikins (Bradley and Lang 1994; Balconi et al. 2015; Bandara et al. In press) and NASA TLX. Fundamentally, this boils down to asking the subject – were you upset, overloaded, angry, sad, etc. This requires a certain amount of self-awareness on the subjects, a fair degree of honesty (Paulhus and Vazire 2009), and a good recollection of how they felt over a period of time without succumbing to recency bias (Sackett and Larson Jr. 1990; Morrison et al. 2014).

Any one of the instruments listed above is considered well-validated *gross* measures of emotional or cognitive state. However, sensors sample anywhere between 1 and 1000 times per second, which means the subject’s state needs to be accurately labeled for each of the intervals. For example, fNIRS focuses on the hemodynamics of the brain, and most devices sample somewhere in the vicinity of 10 Hz, tracking blood flow to regions that changes measurably within a 6–8 s window. Yet, we use a subject’s best estimation of their mental state over a 16–30 s window, hoping that the most recent impression of their mental state does not prompt them to ignore the mode state. Researchers have noted repeatedly that self-reports do not have guaranteed accuracy, with some suggesting that a best practice is to triangulate their observations with other known, validated measures (Liu et al. 2015; Rusnock et al. 2015).

Response based labeling has certain advantages – it can be used to triangulate task-based labeling (see below), or to explore meaningful concepts that are tested using protocols that are known to reliably induce cognitive or emotional responses. An additional advantage of response based labeling is that it may also make it easier to connect the machine learning body of literature to other HCI literature that still relies heavily on self-report measures (Rek et al. 2013; Lottridge 2009; Olson and Kellogg 2014).

3.2 Task-Based Labeling

Task-based labeling involves using tasks that are known or expected to elicit certain mental or emotional states reliably (Ang et al. 2012; An et al. 2013). Task-based labeling is not reliable, even for well-established measures. For example, researchers developed and tested a game that was perceived to have multiple levels of difficulty and thus expected to provoke different levels of engagement. However, during their experiment, two subjects did not notice the difference in difficulty levels and seven did (Girouard et al. 2009). Any attempts to build models using fixed channels as inputs, where the channels are expected to map to specific areas of the brain, faces challenges as well. In some studies, handedness influences cerebral blood flow on certain tasks (Cuzzocreo et al. 2009). Finally, task-based labeling is built on the assumption that participants are engaged in the task.

Task-based labeling has certain advantages, with the most notable being that they avoid the limitations of response-based labeling of conditions. There are two additional benefits of task-based labeling – (1) it allows researchers to more accurately track expected changes in cognitive state because expected changes can be synced to changes in the task; and (2) researchers do not have to interrupt the flow of the experiment to ask the subject to rate his or her experience.

3.3 Performance-Based Labeling

Performance-based labeling involves establishing ground truth by measuring the subject's performance on a specific task. The general chain of assumptions appears to be that (a) the task relies on known cognitive processes, (b) performance on the task requires effort, and (c) performance is correlated with activation and failure is correlated with lack of activation. An example of performance-based labeling can be found in (James et al. 2010), where the researchers estimated cognitive burden generated by learning a visual-motor task by measuring the distance from the cursor to the target on the screen.

Performance-based labeling avoids the pitfalls of self-report measures in that they offer temporal granularity and do not require subjects to estimate their own state. They also avoid some of the limitations of task-based measures, most notably addressing the concern that subjects may or may not be engaged in the task. Performance-based labeling does not address limitations in terms of accurately localizing activation in individuals, although determining the subject's handedness appears to account for a large portion of behavioral lateralization (Lawlor-Savage and Goghari 2014).

4 Conclusion

In this paper we presented a provisional taxonomy for determining ground-truth of emotional or cognitive states in experiments involving the use of machine learning on neurophysiological data. Each approach has strengths and weaknesses, and researchers can either determine those limitations are within the limits of acceptability or employ triangulation procedures to improve their confidence in the measures. We are not arguing that researchers should always use triangulation (although it would be beneficial); instead, we would like to suggest starting a discussion on how it would be possible to estimate the upper boundary of accuracy for the algorithms based on the acknowledgement that the models were trained on data that had low *n* and was only partially accurate.

References

Ang, K.K., Yu, J., Guan, C.: Extracting and selecting discriminative features from high density NIRS-Based BCI for numerical cognition. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–6 (2012). doi:10.1109/IJCNN.2012.6252604

An, J., Lee, J., Ahn, C.: An efficient GP approach to recognizing cognitive tasks from fNIRS neural signals. Sci. China Inf. Sci. **56**(10), 1–7 (2013). doi:10.1007/s11432-013-5001-8

Balconi, M., Grippa, E., Vanutelli, M.E.: What hemodynamic (fNIRS), Electrophysiological (EEG) and autonomic integrated measures can tell us about emotional processing. Brain Cogn. **95**, 67–76 (2015). doi:10.1016/j.bandc.2015.02.001

Bandara, D., Song, S., Hirshfield, L., Velipasalar, S.: A more complete picture of emotion using electrocardiogram and electrodermal activity to complement cognitive data. In: HCI International 2016 Conference Proceedings. Springer, Toronto (In press)

Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)

Cuzzocreo, J.L., Yassa, M.A., Verduzco, G., Honeycutt, N.A., Scott, D.J., Bassett, S.S.: Effect of handedness on fMRI activation in the medial temporal lobe during an auditory verbal memory task. Hum. Brain Mapp. **30**(4), 1271–1278 (2009). doi:10.1002/hbm.20596

Fairclough, S., Gilleade, K.: Advances in Physiological Computing. Springer Science & Business Media, New York (2014)

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., Dehais, F.: Real-time state estimation in a flight simulator using fNIRS. PLoS ONE **10**(3), e0121279 (2015). doi:10.1371/journal.pone.0121279

Girouard, A., Solovey, E.T., Hirshfield, L.M., Chauncey, K., Sassaroli, A., Fantini, S., Jacob, R.J.: Distinguishing difficulty levels with non-invasive brain activity measurements. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 440–452. Springer, Heidelberg (2009)

Girouard, A.: Towards adaptive user interfaces using real time fNIRS. Tufts University, Medford, MA, USA (2010)

Hoskin, R.: The dangers of self-report. Sci. Brainwaves, 3 March 2012. http://www.sciencebrainwaves.com/the-dangers-of-self-report/

James, D.R., et al.: Cognitive burden estimation for visuomotor learning with fNIRS. In: Jiang, T., Navab, N., Pluim, J.P., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 319–326. Springer, Heidelberg (2010)

Lawlor-Savage, L., Goghari, V.M.: Working memory training in schizophrenia and healthy populations. Behav. Sci. **4**(3), 301–319 (2014). doi:10.3390/bs4030301

Liu, N., Cui, X., Bryant, D.M., Glover, G.H., Reiss, A.L.: Inferring deep-brain activity from cortical activity using functional near-infrared spectroscopy. Biomed. Opt. Express. **6**(3), 1074–1089 (2015). doi:10.1364/BOE.6.001074

Lottridge, D.: Evaluating human computer interaction through self-rated emotion. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5727, pp. 860–863. Springer, Heidelberg (2009)

Marx, A.-M., Ehlis, A.-C., Furdea, A., Holtmann, M., Banaschewski, T., Brandeis, D., Rothenberger, A., et al.: Near-Infrared Spectroscopy (NIRS) neurofeedback as a treatment for children with Attention Deficit Hyperactivity Disorder (ADHD)—a pilot study. Front. Hum. Neurosci. **8**, 1038 (2015). doi:10.3389/fnhum.2014.01038

Morrison, A.B., Conway, A.R., Chein, J.M.: Primacy and recency effects as indices of the focus of attention. Front. Hum. Neurosci. **8** (2014). doi:10.3389/fnhum.2014.00006

Naseer, N., Hong, K.-S.: fNirs-based brain-computer interfaces: a review. Front. Hum. Neurosci. **9** (2015). doi:10.3389/fnhum.2015.00003

Noah, J.A., Ono, Y., Nomoto, Y., Shimada, S., Tachibana, A., Zhang, X., Bronner, S., Hirsch, J.: fMRI validation of fNIRS measurements during a naturalistic task. J.Visualized Exp. 100 (2015). doi:10.3791/52116

Olson, J.S., Kellogg, W.A.: Ways of Knowing in HCI. Springer Science & Business, New York (2014)

Paulhus, D.L., Vazire, S.: Thse self-report method. In: Robins, R.W., Chris Fraley, R., Krueger, R.F. (eds.) Handbook of Research Methods in Personality Psychology, pp. 224–239. Guilford Press, New York (2009)

Rek, M., Romero, N., van Boeijen, A.: Motivation to self-report: capturing user experiences in field studies. In: Collazos, C., Liborio, A., Rusu, C. (eds.) CLIHC 2013. LNCS, vol. 8278, pp. 111–114. Springer, Heidelberg (2013). doi:http://link.springer.com/chapter/10.1007/978-3-319-03068-5_19

Rusnock, C., Borghetti, B., McQuaid, I.: Objective-analytical measures of workload – the third pillar of workload triangulation? In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2015. LNCS, vol. 9183, pp. 124–135. Springer, Heidelberg (2015)

Sackett, P.R, Larson Jr., J.R.: Research strategies and tactics in industrial and organizational psychology. Dunnette, M.D., Hough, L.M. (eds.) Handbook of Industrial and Organizational Psychology, vol. 1, 2nd edn. Consulting Psychologists Press, Palo Alto (1990)