IMPROVING MACHINE LEARNING PREDICTION OF CONSTRUCTS: MENTAL FATIGUE

Vojtěch Formánek, & Vojtěch Juřík
[1]*Department of Psychology, Faculty of Arts, Masaryk University (Czech Republic)*

Abstract

Mental fatigue is a psychophysiological state that plays an important role in various domains of human machine interaction where it may increase risk of injury or accidents. To prevent threats to life and property, novel techniques combining psychological and computational approaches are needed and thus explored. Previous research has focused on training machine learning (ML) models on different types of fatigue input data and experiment settings, and recently on the generalizability of the models. However, current ML development struggles with various issues such as unclear analysis of what the model is actually learning. When trained on data that are only partially correctly labeled, it can learn artifacts of the dataset construction instead of the construct state. Psychometric measures that are used to label data have usually imperfect/questionable reliability, thus even if administered correctly may label some data incorrectly. The most widely used method for labeling mental fatigue states are subjective scales, which also possess limitations on construct validity. In this contribution, an iterative procedure to improve both reliability and validity of the labeling based on generalizability theory is proposed. This labeling procedure is constructed from components already present in the dataset and relevant to the construct being predicted. In the case of mental fatigue, a subjective scale, performance decrease and environmental reference extracted in 7 datasets collected on different sites is used, with several methods used to induce fatigue, all with heart rate variability as input data. The quality of combinations and levels of the label is assessed by analyzing unwanted variances and by using an equivalent of reliability from generalizability theory. Applying this procedure, components can be added to a label and created labels can be directly compared. Considering the iterative nature of this process, labels can be dynamically adjusted based on added new data. The whole procedure adds flexibility to dataset design, allowing for easier integration of datasets, even those that were not originally intended for ML. As a result, we enhance increasing variability and amount of data that is available for researchers, promoting its use beyond the ML-based mental fatigue prediction.

Keywords: *Machine learning, fatigue, generalizability, reliability*.

1. Introduction

Machine learning (ML) offers great opportunities for various applied domains, including psychology. Specifically, mental fatigue, a psychophysiological state playing a crucial role, e.g., in efforts to prevent human and property threats, represents a promising area of ML research. When training a model to classify construct states, we want it to learn a structure from the input data related to the construct state. A label gives the model information about the input data's state. For example, suppose we want it to identify a given participant's fatigue from his physiological data. To train correctly, it needs information about the state in which it should classify the data. Thus, when predicting construct states, integration of psychometrics and ML is necessary because the measures typically used to assess these constructs were not intended for ML. Therefore, they should be examined directly before training the model for better transparency and control of what the ML model is learning. Various methods are used to measure constructs, typically through a test and environmental manipulation, the changes in the construct states are hypothesized and observed through the changes in the test scores. In essence, the score is used to estimate the construct state, and the quality of this estimation is described by two concepts known to psychology - reliability and validity. Both can be used more generally to assess the quality of any measure.

Generally, fatigue as a construct suffers from two significant issues in terms of validity. There is a long-standing lack of consensus on the preferable measurement methods of fatigue and, more importantly, on the definition of fatigue *per se* (Bartley et al., 1947; Phillips, 2015). Therefore, it gets mistaken for related constructs, such as sleepiness (Shen et al., 2006) or tiredness and exertion (Phillips, 2015).

Regardless, there are several options available for its measurement. According to Martins (2021), considering ML modeling, the most prominent reference measures are the Karolinska Sleepiness Scale (KSS; Åkerstedt & Gillberg, 1990) and Borg's Rating of Perceived Exertion (RPE; Borg, 1998), both one-dimensional subjective scales. Alongside the abovementioned issues, fatigue is usually identified as a post-experiment state (Matuz, 2022), essentially using current exertion to label the data.

In Classical Test Theory (CTT), reliability is defined as a correlation between the latent, noiseless test score and the measured test score (Revelle et al., 2019). This means that the noise of the label is equal to the shared variances of the latent and measured score, which is a square of the correlation. Even for tests with high reliability while at the same time assuming perfect administration conditions, the amount of noise introduced by the label can be devastating to the ML model's performance, including also those typically considered robust against label noise, such as Random Forests (Reis, Baron, & Shahaf, 2018). For a reliability of 0.9, which is usually considered high, the number of wrongly classified labels is 19 % ($1-0.9^2$). This number increases exponentially as reliability decreases. For example, for a reliability of 0.8, which is still considered relatively high, the amount goes to 36 %. Typically, techniques in ML require larger datasets than those currently publicly available for mental fatigue prediction since the largest available dataset includes data from 85 participants (Matuz, 2022). Other techniques require a subset of correct labels (Northcutt., Jiang & Chuang, 2021) or other methods of assessing the probability of using an incorrect label (Reis et al., 2018).

Fatigue is an inseparable part of mental fatigue. Phillips (2015) defines it as "a suboptimal psychophysiological condition caused by exertion." According to him, the degree and dimension depend on the dynamics and context of exertion, described by the value and meaning of performance to the individual and environmental conditions. Fatigue changes strategies or resource use such that original mental processing or physical activity levels are maintained or reduced (Phillips, 2015). Further, mental from physical fatigue can be differentiated as a decrease in mental processing as opposed to physical activity. Exertion in all the below-mentioned datasets is induced by cognitive tasks that require sustained attention.

Hood (2009) integrates two existing definitions of validity; broadly, it refers to the adequacy of inferences made based on the measure (from Messick, 1989), and a measure is valid if the attribute exists and variation in it causes variation in the measure (from Borsboom et al., 2004). The question of whether subjective experience, which is what is primarily measured by subjective scales, fully saturates the mental fatigue construct has been raised (Phillips, 2015) and then partially answered by adding other dimensions to the definition of fatigue. The proposed approach adds the measures for these dimensions and evaluates this composite measure. The former is done to increase validity, estimate reliability, better describe the data for the model, and help us understand what it is learning. Here, generalizability theory (GT; Brennan, 2001) can be used to resolve these constraints since GT, compared to CTT, focuses on analyzing various sources of unwanted error variance in the data.

2. Methods

Regarding the abovementioned, to follow the goal of effective ML training of fatigue construct, we found it advantageous to include multiple reference variables, i.e., multiple dimensions of fatigue in the context of generalizability theory. To demonstrate this, in the framework of this study, the existing datasets (4 in total; see below for more detail) containing fatigue scores collected from various psychological experiments were analyzed. Based on this, we included four dimensions of target measures: absolute performance score, absolute subjective feeling, and relative changes of these two variables from the previous measurement. These are reconstructed from already collected data; we're creating a four-dimensional measure.

2.1. Datasets

Due to technical problems, we eventually used only 4 out of the 7 datasets we had available, one unpublished (Pešán et al.; 2023) and three published by Matuz (2024). The former was designed for stress prediction; however, it does use custom cognitive tasks to exert participants and subjective load to assess the participants' state (Pešán et al., 2024). Matuz uses a Visual Analog Scale (VAS), and the difference between the three datasets they composed is in the tasks used: task-switching, 2-back, and the Stroop experiment. All of them are attention-based and exert cognitive capacity, and this way induces mental fatigue. Matuz does not include the results of these tasks in the available dataset; thus, we reconstruct them using K-Nearest Neighbors (KNN). For further analysis, we assume that the scores for the two categories are drawn from identical distributions.

Given that the latter three datasets provide only two subjective measures sampled at different times, we shall consider only two measurement phases, pre- and post-experiment-induced exertion.

Absolute performance is taken as a numerical score extracted from the above tasks, normalized from 0 to 1, and then combined with the other datasets. Subjective experience is, in this case, only normalized. The relative changes of the phases are taken as the difference of the scores; we invert performance change since it correlates negatively with increased scores of subjective experience, so that a low value is related to no/low mental fatigue, and the higher the value is, the more exhausted a person is. Three bins with different thresholds for each variable are used to simplify label creation.

2.2. Reliability estimation based on generalizability theory

Evaluating the reliability of this composite measure using the traditional approaches is almost impossible since the data have already been collected without this type of analysis in mind. An alternative is to use GT, which, regarding the ML domain, overcomes CTT in its feature of analyzing various sources of unwanted error variance in the data (Brennan, 2001), as discussed above. We follow the general outline of Briesch et al. (2014). Firstly, the universe of admissible observations includes all possible reference variables, datasets, phases, and persons. The last is also the object of measurement. Facets represent the error sources in the design; these include all of the variables in the universe, excluding persons. The design is then (person: dataset) × reference variable, using nested facets and random effects. GT then estimates the variances of these facets using linear models with mixed effects and uses two coefficients, dependability (Φ) and generalizability ρ^2), these are defined as:

$$\Phi_p = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{abs}^2}, \quad \rho_p^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{rel}^2}, \quad \sigma_{abs}^2 = \frac{\sigma_{gr}^2}{n_r} + \frac{\sigma_e^2}{n_r}$$

since Φ is more conservative (Briesch, 2014), we use only it when analyzing to avoid confusion. A guide on GT in Briesch et al. (2014), in-depth description of alternative designs in Webb (1991), we analyzed the data in R (described in Huebner, 2019) and the variances using the gtheory package and formula: *fatigue ~ (1|reference_variable) + (1|dataset/person) + (1|reference_variable: dataset)*.

2.3. Mental fatigue prediction

Finally, the label is constructed by summing the values of the reference variables and distributing them into three bins: the first contains values indicating no/low mental fatigue [0-5], the second indicating medium mental fatigue [6-10] and the last indicating high mental fatigue [11-12]. All the datasets contain heart rate data from which we preprocessed and extracted indices. For Pešán et al.'s (2023a) dataset, we follow the process from Matuz (2022) to make the input data as similar as possible. Datasets were processed in Python; for data manipulation, we used Pandas (McKinney, 2010) and Numpy libraries (Harris, Millman, & van der Walt, 2020), after loading filtering and feature extractions were done using neurokit2 (Makowski et al., 2021).

However, we were not dealing with the distribution shift of the extracted indices, which roughly refers to systematic differences in collected data. This is likely because data was collected at different sites using different devices. Thus, to evaluate the label and train the models, we only use our dataset since it is the only one from which we extracted all the data and does not require dealing with the distribution shift. Three models were trained: Random Forests, Support Vector Machines (SVM) and K-Nearest Neighbors, and a baseline model that predicts the most frequent class. Models are implemented in scikit-learn (Pedregosa et al., 2011), which was also used to impute missing data. The models are evaluated using accuracy and 10-fold cross-validation with 20 repeats while keeping the original distribution of the classes.

3. Results

In terms of gender, the datasets are imbalanced, in Pešán et al.'s (2023) dataset, there were 18 women (11.5 % of the total) and 54 men (39.7 % of the total); in Matuz (2022), 33 women (20.9% of total) and 52 men (32.9 % of total). All variables are ordinal and found non-normal upon inspection using Q-Q plots. For this reason, we used the Mann-Whitney U test for independent samples to test the null hypothesis that VAS and subjective load come from the same distribution on the four datasets.

The difference between the subjective experience (SE) measures in all the datasets was not found to be significant (U=8712, p=0.32). Similarly, the difference in SE measures before exerting participants was also not significant (U=2568, p=0.32), but after exerting them, it was significant (U=1750, p<0.01), demonstrating a difference between the SE measures, either a difference in the exertion level, the measures or simply datasets. For performance, none of the test results are significant, without differentiating exertion (U=9759, p=0.54), on the pre-exertion measurement (U=2412, p=0.76) nor post (U=2724, p=0.09).

3.1. Reliability estimation and model evaluation

The dependability index for the composite measure pre-exertion is rather low (Φ = 0.21); no variance is caused by datasets or interaction with reference variables. Relatively slight variance caused by the object of measurement (6%, σ_p^2 = 0.01); this is in part caused by how we dealt with relative changes since we always substituted a zero to imitate no change at the beginning; this is supported by the fact that 30.8% (σ_s^2 = 0.07) of the total variance is caused by the reference variables. The residual variances cover the largest proportion (63.1%, σ_e^2 = 0.15). In the broader context, we would expect the first measurement to be inconsistent because the participants come to the experiment from different environments and have various levels of mental fatigue.

The dependability post-exertion is much higher (Φ = 0.68), with negligible influence caused by the datasets (6.2%, σ_d^2 = 0.019), reference variables (12.6%, σ_r^2 = 0.039) and their interaction (2.2%, $\sigma_{r:d}^2$ = 0.007). The object of measurement accounts for a larger share of the total variance (30.2%, σ_p^2 = 0.094); the residuals still account for almost half of the variance (48.8%, σ_e^2 = 0.152).

Only RF beat the baseline by about 0.05 with a resulting accuracy of 0.44, both above chance (0.33) and picking the most frequent class (0.39). This indicates that the model can learn this label, but not satisfactorily. However, note that the models were not fine-tuned; thus, the resulting score could be higher.

4. Discussion

In this contribution, we made the first step for a robust approach to combining 4 datasets from 2 different sites, continued in the research to increase the generalizability of machine learning models trained to predict mental fatigue, and outlined the approach so that it can be used outside both mental fatigue and ML. The results show that the reference variables after exerting participants are much more reliable, although their reliability is still low. The first phase also requires much more care and consideration in future research.

Though neither the reliability nor model performance overall are high at this stage, the strength of this approach lies in its flexibility and many possible points of improvement. First is in terms of validity of the reference variable - a multitude of other measures can be added, especially physiological measures that are not directly related to the input data, such as indices extracted from electroencephalography (EEG). Before estimating dependability, transformations can be applied to the measure's result to increase reliability and validity. However, it should always be considered a priority since a transformation that puts all reference measures to the same constant will result in high reliability. Also, for simplicity, we have assumed that along a given dimension, all the datasets use the same reference measure, for example, subjective experience. If this is not the case, the measures can be transformed before being combined.

Also, we used a simple way of transforming reference variables into a label. Generally, any transformation can be applied; a simple improvement is weighing the individual scales. Still, both should be based on previous research and considered in terms of their validity. Ideally, the transformations should label the data as best as possible so that the ML model can distinguish the individual examples more clearly.

The most challenging problem when combining datasets is the distribution shift. We found that this is the case by visually analyzing and comparing the distributions. Several possible causes are present from experiment conditions, devices that collected the ECG data, preprocessing, and feature-extracting strategies. Still, with care, these changes can be reflected in how the reference measures are constructed and combined. Most importantly, their reliability can be evaluated using GT.

Acknowledgments

The paper was supported by the project MUNI/A/1519/2023 - Possibilities of applying knowledge of basic psychological research. The study was also supported by the research infrastructure GREY Lab held at the Department of Psychology, Faculty of Arts, Masaryk University, Brno.

References

Adão Martins, N. R., Annaheim, S., Spengler, C. M., & Rossi, R. M. (2021). Fatigue monitoring through wearables: a state-of-the-art review. *Frontiers in physiology*, *12*, 790292.

Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International journal of neuroscience*, *52*(1-2), 29-37.

Bartley, S. H., & Chute, E. (1947). Various views on fatigue.

Borg, G. (1998). *Borg's perceived exertion and pain scales*. Human kinetics.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*(4), 1061.

Brennan, R. L., & Brennan, R. L. (2001). Variability of statistics in generalizability theory. *Generalizability theory*, 179-213.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of school psychology*, *52*(1), 13-35.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. (2020). *Array programming with NumPy*. Nature 585, 357–362.

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology*, *19*(4), 451-473.

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, *24*(1), 5.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4), 1689-1696. https://doi.org/10.3758/s13428-020-01516-y

Matuz, A., Van der Linden, D., Darnai, G., & Csathó, Á. (2022). Machine learning models in Heart Rate Variability based mental fatigue prediction: training on heterogeneous data to obtain robust models.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *Ets research report series*, *1993*(2), i-18.

McKinney, W. (2010). Data structures for statistical computing in Python. In *SciPy*.

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373-1411.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pešán, J., Juřík, V., Kolářová, J., Chudý, P. & Černocký, J. (2024). BESST: An Experimental Protocol for Collecting Stress Data for Machine Learning. [Manuscript submitted for publication]. Brno University of Technology

Pešán et al. (2023). Speech production under stress for Machine Learning: multimodal dataset of 79 classes and 8 signals. [Manuscript submitted for publication]. Brno University of Technology

Phillips, R. O. (2015). A review of definitions of fatigue–And a step towards a whole definition. *Transportation research part F: traffic psychology and behaviour*, *29*, 48-56.

Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, *157*(1), 16.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological assessment*, *31*(12), 1395.

Shen, J., Barbera, J., Colin M., (2006). Distinguishing sleepiness and fatigue: focus on definition and measurement. Sleep medicine reviews. Vol. 10, no. 1, pp. 63–76.

Webb, N. M. (1991). Generalizability theory: A primer.