Predicting EEG Sample Size Required for Classification Calibration

Zijing Mao[1], Tzyy-Ping Jung[2], Chin-Teng Lin[3], and Yufei Huang[1](✉)

[1] Department of Electrical and Computer Engineering,
University of Texas, San Antonio, TX, USA
mzjl68@hotmail.com, yufei.huang@utsa.edu

[2] Institute for Neural Computation,
University of California, San Diego, CA, USA
jung@sccn.ucsd.edu

[3] Brain Research Center, National Chiao Tung University,
Hsinchu, Taiwan
ctlin@mail.nctu.edu.tw

Abstract. This study considers an important problem of predicting required calibration sample size for electroencephalogram (EEG)-based classification in brain computer interaction (BCI). We propose an adaptive algorithm based on learning curve fitting to learn the relationship between sample size and classification performance for each individual subject. The algorithm can always provide the predicted result in advance of reaching the baseline performance with an average error of 17.4 %. By comparing the learning curve of different classifiers, the algorithm can also recommend the best classifier for a BCI application. The algorithm also learns a sample size upper bound from the prior datasets and uses it to detect subject outliers that potentially need excessive amount of calibration data. The algorithm is applied to three EEG-based BCI datasets to demonstrate its utility and efficacy. A Matlab package with GUI is also developed and available for downloading at https://github.com/ZijingMao/LearningCurveFittingForSampleSizePrediction. Since few algorithms are yet available to predict performance for BCIs, our algorithm will be an important tool for real-life BCI applications.

Keywords: Sample size prediction · Calibration · Brain computer interface · EEG · Rapid serial visual presentation · Driver's fatigue

1 Introduction

A brain computer interaction (BCI) system allows interactions between human and an external device through monitoring brain signals [1]. EEG-based BCIs have become increasingly popular in the past decade, finding real-life applications from controlling wheel chairs to monitoring human performance [2]. Most of the BCI systems require a calibration stage, where training samples are collected to build a classification model for event detection from brain signals. The current practice of BCIs relies on collecting an excessively large amount of calibration data to ensure that a robust classifier can be built. Such practice has become a bottleneck for the BCI applications in real-world

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part I, LNAI 9743, pp. 57–68, 2016.
DOI: 10.1007/978-3-319-39955-3_6

settings because such practice prolongs BCI training time and deteriorates user performance due to induced fatigue on users. Moreover, due to individual differences in brain responses, calibration needs to be adapted for each individual. Related efforts have been made to take advantage of machine learning (ML) algorithms such as active learning [3] and transfer learning [4, 5] by borrowing from existing data from the same or other subjects to reduce the calibration samples as much as possible. However, as long as there is a need for collecting calibration samples, determining an appropriate calibration sample size for each individual before the training is an important issue to be tackled. In fact, integrating sample size prediction together with transfer learning in the calibration stage should be a favorable practice.

Despite its importance, the problem for predicting calibration sample size for BCIs has not received much attention in the past. However, the problem of sample size estimation (SSE) [6] has been studied in many other fields for different purposes. The existing work stems mostly from three main types of methods. The first type is the power analysis for sample size calculation [7], a method that is widely applied in biostatistics, bioinformatics, and clinical research [7–9]. Power analysis requires information about effect size, significance level, and power of the underlying hypothesis testing to predict the sample size; sophisticated tools [9] have been developed for this analysis. However, power analysis concerns more on the statistical significance rather than the classification performance as in BCIs. The second type of methods treats SSE as an optimization problem and defines specific optimization functions to balance the cost and benefit of using a sample size [10]. However, these optimization-based SSE methods require knowledge to define cost and benefit in the same domain such that they can be compared and thus optimized; this knowledge is difficult to obtain in many applications including BCIs. For instance, while it is possible to assess the cost of collecting samples in BCIs in terms of money or time, it is nevertheless difficult to assess the benefit in performance improvement in terms of cost or time. The last type of methods is the learning curve fitting, which fits a regression model to the observed sample sizes and performances to capture the relationship between performance and sample size. Since 1936, learning-curve fitting has been studied and applied in many industrial fields [11]. One of the most widely used fitting model is the inverse power law [11], by the intuitive thought of more samples always improve the performance but improvements decay gradually. Because of its data-driven nature and ease of implementation, we apply it for the prediction of calibration sample sizes for BCIs.

We propose a novel adaptive algorithm for EEG calibration-sample-size prediction. The algorithm has several unique features tailored for BCI tasks. First, the algorithm utilizes the prior datasets commonly available in BCIs to suggest a baseline performance and to derive a population-wide sample size upper bound. Second, it adaptively fits the learning curve between performance and sample size for each individual and makes the prediction of calibration sample size when a satisfactory fitting confidence level is reached. Third, it also provides a way to identify subject outliers that potentially need excessive amount of calibration data. Fourthly, it can be used to select the best classifier for BCIs. We evaluated the algorithm and demonstrated its efficacy on three different BCI datasets. A Matlab package with GUI is also developed and released to facilitate the application of the proposed methods (https://github.com/ZijingMao/LearningCurveFittingForSampleSizePrediction).

The remainder of the paper is organized as follows. Section 2 discusses the proposed algorithm in details. Testing results are reported in Sect. 3. Conclusions are drawn in Sect. 4.

2 Materials and Methods

2.1 Experiments and Data

Data from three BCI experiments are used in this study to test the proposed algorithm. The experiments include two image Rapid Serial Visual Presentation (RSVP) [12] experiments and one simulated driving experiments for driver performance study. The RSVP experiments are Static Motion (D1) and the Cognitive Technology Threat Warning System (CT2WS or D2) [13, 14]. The static motion RSVP experiments include the presentations of color target images of enemy soldiers/combatants versus the background non-target images of village street scenes. The CT2WS experiment includes presentations of gray scale images, where target images include moving people and vehicle animations, whereas the non-targets are other types of animations such as plants or buildings. Each subject performed four sessions in static motion and only one session in CT2WS, where each session lasted for about 15 min. For both experiments, the images were presented at 2 Hz (one image presented every 500 ms) and brain signals were recorded with 64-channel Biosemi EEG systems at a sampling rate of 512 Hz. There were a total of 16 and 15 subjects in the static motion and CT2WS experiment, respectively. The simulated driving dataset (D3) includes EEG samples from 17 subjects, each performed a lane-keeping driving task in a virtual reality interactive driving platform with a 3-D highway scene [15]. Perturbations to the car were introduced into driving path every 8 to 12 s and driver's reaction time and the amount of the lane deviation was measured to assess the degree of driver's drowsiness. Each experiment lasted one and half hours during which EEG signals were measured from 30 electrodes. The reaction time (RT) is defined as the time between the onset of the lane perturbation and the moment when the subject starts steering the car. RT is used to define the drowsy or alert state of the driver. Particularly, when the reaction time is ≤ 0.7s, the driver is considered as alert, whereas when the reaction time is ≥ 2.1s, the driver is considered as drowsy.

2.2 Data Preprocessing

EEG data from three experiments were subject to the similar preprocessing steps. Particularly, the raw EEG data were first bandpass-filtered with a bandwidth ranging from 0.1–50 Hz in order to remove DC noise and electrical artifacts. Down-sampling was performed next to reduce the sampling rate from the original 512 Hz to 128 Hz, which is the maximum down-sampled frequency that does not produce aliasing at the high-passed frequency. Then, by following [16], one-second epochs of the EEG samples after each image onset were extracted for all the subjects to be used as data for calibration and prediction. In the end, about 13,500 epochs from Static Motion ($\sim$1000 epochs per subject) and about 10,400 epochs from CT2WS ($\sim$700 epochs per subject) were obtained. For the driving data, we used one-second epochs before the onsets of the perturbations as data for predicting the “drowsy” or “alert” state of the drivers. There is a total of 2,796 (764 drowsy and 2,032 alert) epochs from the 17 subjects. Because the sampling rate is 250 Hz, the dimension of one-second EEG epoch is 250 × 30 = 7500. Afterwards, normalization was applied to all the epochs. Each (channel × time) pair in the calibration set was normalized across epochs by z-score normalization. The test sets were then z-score normalized according to the calibration set mean and standard deviation. The goal for RSVP classification is to predict if the subject sees a target image based on the epoch data while for driving performance classification, the goal is to predict if the subject has a slow reaction time.

2.3 The Proposed Scheme for Calibration Sample Size Prediction

The goal of calibration sample size prediction is to suggest an appropriate sample size for calibrating the classification algorithm for a new subject. We consider a common scenario in BCIs, where the prior datasets collected from other subjects performing the same task are available and therefore a baseline performances P_B (e.g. P_B = 0.9 Area under ROC or AUC) for satisfactory event classification is learned. Intuitively, an appropriate sample size is the one needed for a classifier to reach the baseline performance, or the baseline sample size S_B as we will refer to next and we hope to predict S_B, denoted as $\widehat{S_B}$ for an individual by collecting only a small number of calibration samples from the subject. To this end, we propose an adaptive algorithm, where at the mth iteration, M new samples are collected and an intermedium prediction and its confidence are calculated using all the samples collected thus far. The final prediction is reached when the prediction confidence falls within a tolerate threshold (e.g. 95 % significance level). At the mth iteration, to make a prediction, a learning curve is first fitted to the performance of a classifier. A learning curve characterizes the classification performance (p_{Az}) as a function of calibration sample size s and as in [8], can be represented using an inverse power law (IPL) model [11]

$$p(s) = f(s; a, b, c) = a \times s^b + c, \tag{1}$$

where a, b and c, are the model parameters that represent the decay rate, learning rate, and bias, respectively. The goal of fitting is to estimate the parameters using the classification performances obtained at all m iterations. To this end, the non-linear least square method is applied in this work and a 95 % confidence interval $I(s)$ of the fitting for the sample size s is also reported. An illustration of this process is shown in Fig. 2. As can be seen, using the learning curve, the sample size s_A can be predicted from (1) by setting $p(s) = P_B$. Then, $I(s_A)$, the 95 % confidence interval at $s = s_A$, is compared with a predefined tolerance level T_s. We define T_s by calculating the ratio between curve fitting confidence interval bound and s_A. For example, if we set T_s = 0.02, it means the range of confidence interval is 2 % of s_A. If $I(s_A) < T_s$, then s_A is reported as the predicted baseline sample size; otherwise additional M samples will be collected and one additional iteration will be performed.

For some subjects, S_B can be excessively large and it might not be prudent to collect such large samples given limited resources. To determine if S_B is too large, we resort to the prior datasets. Particularly, we bootstrap the dataset 100 times and for each bootstrapped data, we perform cross-validation to determine the baseline sample size S_B. Then, we counted the histogram of all the bootstrapped baseline sample sizes to generate the population-wide distribution of S_B. Based on this distribution, we estimated a population-wide sample size upper bound S_β, as $P(S_B > S_\beta) \leq 0.05$. Given that the prior dataset is large enough to capture the data distribution of the subject population, S_β can be interpreted as a sample size upper bound such that only 5 % of subjects require more samples to reach the baseline performance P_B. Therefore, the predicted baseline sample size $\widehat{S_B}$ for the subject of interest can be determined to be excessively large if $\widehat{S_B} > S_\beta$. In this case, we recommend $\widehat{S_B} = S_\beta$ if there is prior dataset for performing transfer learning. Otherwise, we suggest excluding this subject from this task. Taking together, we report $\widehat{S_B}$ as the predicted calibration sample size if $\widehat{S_B} < S_\beta$; otherwise, we suggest to collect S_β calibration samples and then apply transfer learning (TL) algorithms to improve the classification performance to P_B. The procedure of the algorithm is summarized as follows.

```
Initialize the baseline performance and sample size: $P_B$ and $S_\beta$
Initialize sample size increment $M$ and initial sample size $s_c$
Initialize the tolerance level
While $S_\beta > s_c$ do
    ◆ Obtain calibration performance $p_{Az}(s_c)$ based on $s_c$
    ◆ Based on all obtained $p_{Az}$s, fit learning curve $p(s)$
    ◆ Estimate the baseline sample size
    ◆ Estimate 95% CI, $I(s_A)$
  If $I(s_A) < T_s$ do
        ◆ Predicted baseline sample size $\widehat{S_B} \leftarrow s_A$
      If $p_{Az}(s_c) > P_B$ do
        Break
      End if
    End if
    ◆ Current sample size $s_c \leftarrow s_c + M$
End while
If $\widehat{S_B} > S_\beta$ or $\widehat{S_B}$ does not exist do
    ◆ Set predicted baseline sample $\widehat{S_B} = S_\beta$;
      or exclude the current subject
Else
    ◆ Report predicted baseline sample $\widehat{S_B}$

End if
```

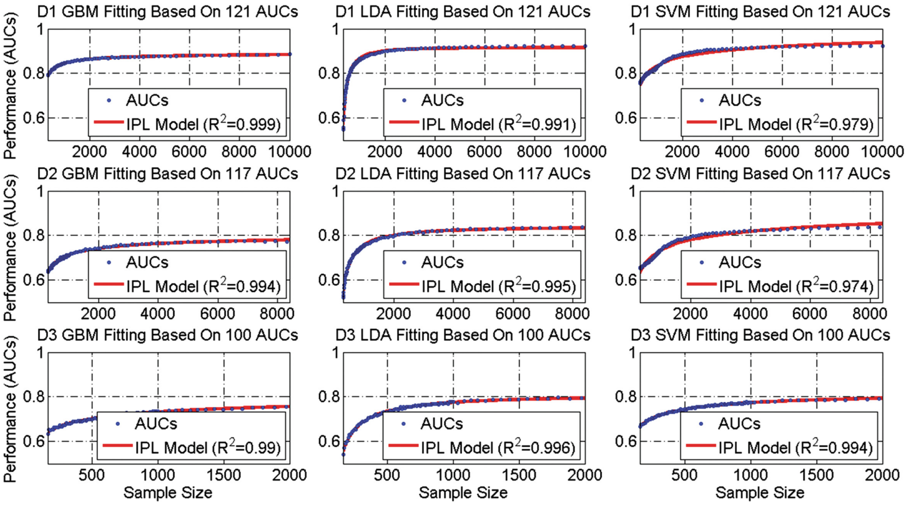

Fig. 1. Learning curve fitting of 3 datasets for 3 classification methods. The blue dots are AUCs obtained by calibration with an increasing size of EEG samples. These AUCs were obtained at sample sizes linear-spaced below 1,000 with a step size of 10 and log-spaced above 1,000 with a step size that amounts to 50 logarithmically even-spaced points between 10^3 and 10^4. (Color figure online)

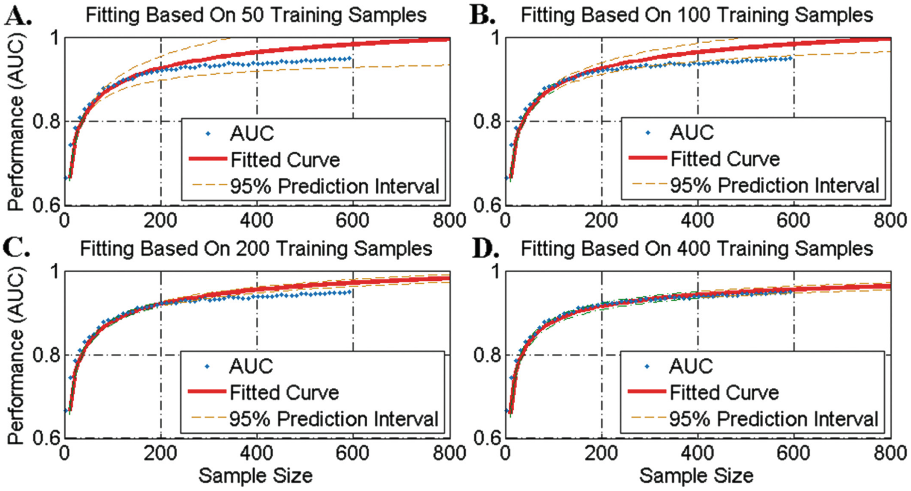

Fig. 2. An illustration of adaptive learning curve fitting. The blue dots are AUCs obtained by calibration with an increasing size of EEG samples. The figures show the fitting results of using different calibration sample sizes, where **A**, **B**, **C**, **D** used the 50, 100, 200, 400 calibration samples (which means using the first 5, 10, 20 and 40 blue dots) to fit the learning curve (red line) respectively. The data come from subject 1 of Static Motion RSVP (D1). (Color figure online)

We can also apply this algorithm to select the best classifier for the BCI task. Specifically, we predict the sample sizes for all candidate classifiers and the one that is associated with the smallest predicted sample size is selected.

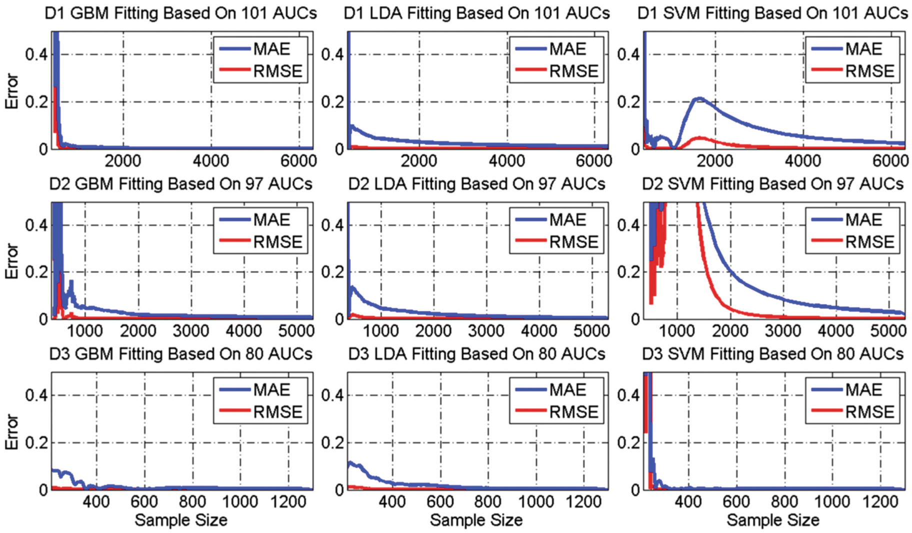

Fig. 3. Mean absolute error (MAE, blue line) and root mean squared error (RMSE, red line) as a function of calibration sample size. The horizontal axis is the sample size used for calibration and obtaining the baseline AUCs, and the vertical axis is the value of MAE and RMSE. (Color figure online)

3 Results

This section demonstrates the performance and utility of the proposed algorithm of calibration sample size prediction using 3 BCI datasets, as described in Sect. 2; and considered 3 classification algorithms including gradient boosting method (GBM) [17], linear discriminant analysis (LDA) [18] and support vector machines (SVM) [18]. As a result, we have nine different combinations for performance evaluation.

3.1 Learning Curve Fitting

We first examined the performance of learning curve fitting, where, for each dataset, we merged samples from all subjects and fitted the inverse power law model for each of the three classifiers, separately. Sample size increased from 300 to 10,000 in D1, from 300 to 8,500 in D2, and from 150 to 2,000 in D3, respectively and at a particular sample size, bootstrapping was performed to calculate the classification ROC of AUCs for each classifier. Figure 1 shows the results of nine fitted learning curves. We also calculated the R^2 statistic as a measurement of the goodness of fit (GoF). The R^2 is denoted by:

$$R^2 = \frac{\sum_n \left(f\left(s_n^{fit}; a, b, c\right) - p_n^{fit}\right)^2}{\sum_n \left(p_n^{fit} - \overline{p_{fit}}\right)^2}, \tag{2}$$

where s_n^{fit} and p_n^{fit} represent the nth fitting sample size and the corresponding ROC of AUCs, the denominator is called the total sum of squares, where $\overline{p_{fit}}$ is the mean of classification performance, and the numerator is called the regression sum of squares. As shown in Fig. 1, R^2 scores are high for all the fitting (> 97 %), indicating that the IPL can model the relationship between performance and sample size.

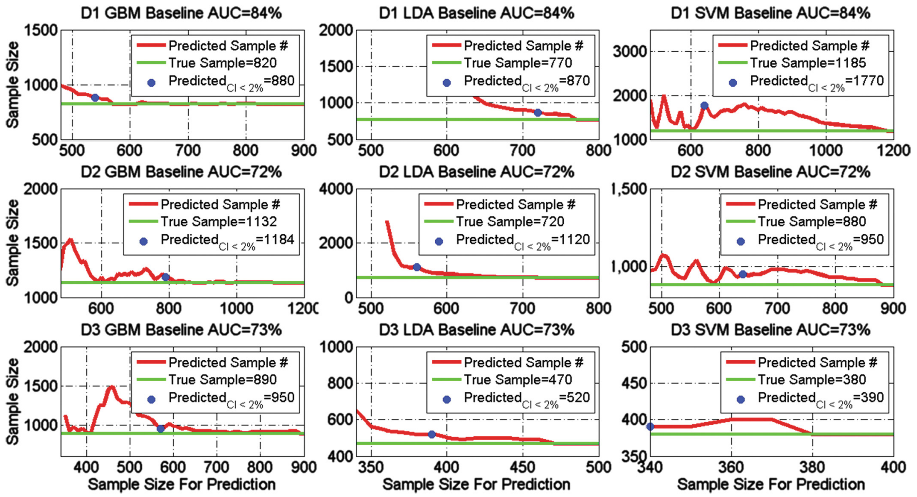

Fig. 4. Adaptive prediction of baseline sample size. The horizontal axis is the sample size used for calibration and obtaining a baseline AUCs. The vertical axis is the sample size predicted by learning curve in order to reach the baseline AUC. The green and red line indicates the true (S_B) and predicted ($\widehat{S_B}$) calibration sample size for the baseline AUC respectively. The blue dot represents the calibration sample size that can reach < 2 % of confidence interval for the learning curve fitting. (Color figure online)

Next, we examined the adaptive fitting of the learning curve and its ability to predict future performance as described above. Figure 2 shows an example of the adaptive fitting and prediction results for Subject 1 in static motion RSVP (D1). In Fig. 2A, the curve was fitted using first 5 AUC points (trained by 50 EEG samples). The predicted curve (red line) deviates slightly from the true AUCs after the 5th point (blue dots) and as expected, both deviations and prediction confidence intervals grow larger as we move further into a larger sample size. However, the true AUCs do fall in between the 95 % confidence interval (orange lines) consistently. As we increased the number of fitting points from 5 to 40 (trained by 50 to 400 EEG samples), the predicted curve became increasingly similar to the behavior of the true AUCs and at the same time the 95 % confidence interval grew much narrower (Fig. 2A–D). At 40 points, the

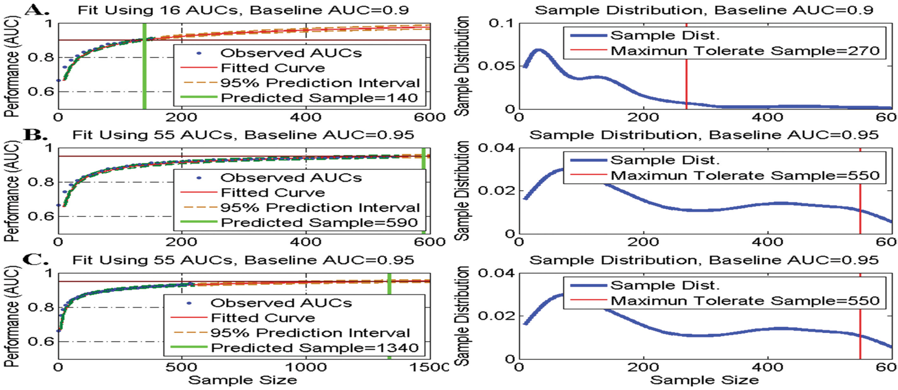

Fig. 5. Illustration of three scenarios that result in different sample size prediction. **A** and **B** are results from subject 1, D1, where the baseline AUCs P_B were set as 0.9 and 0.95, respectively. **C** shows the results from subject 2, D1 where $P_B = 0.95$. The blue dots are AUCs obtained by calibration with an increasing size of EEG samples. The red line is the fitted learning curve. The green vertical line is predicted sample size obtained from fitted learning curve in order to reach P_B. The figures in the right columns depict the distribution of calibration sample size estimated from the prior dataset in order to reach a given P_B. The blue line indicates the sample size distribution and the red line indicates the maximum tolerate sample size for calibration, calculated by the maximum 5 % of the sample size distribution. (Color figure online)

predicted curve closely resembled the true learning curve with very high confidence (Fig. 2D). To systematically evaluate the prediction, we use prediction mean absolute error (MAE) and root mean squared error (RMSE) for each dataset, as in [8], which are defined as

$$MAE = \frac{1}{N}\sum_{n}\left|f\left(s_n^{pred}; a, b, c\right) - p_n^{pred}\right|, \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n}\left(f\left(s_n^{Pred}; a, b, c\right) - p_n^{pred}\right)^2}, \tag{4}$$

where s_n^{Pred} and p_n^{Pred} are n^{th} sample size and corresponding AUC obtained by calibration with n EEG data samples. Figure 3 shows MAE and RMSE as a function of fitting sample size for all three datasets. For each dataset, AUCs associated with 20 largest sample sizes were retained and used for evaluating the prediction MAE and RMSE for learning curve fitted by an increasing sample size. Specifically, D1, D2 and D3 fitted the learning curve using calibration samples range from 300 to 6000, 300 to 5000 and 150 to 1300 with respect. As expected, both RMSE and MAE decrease and became very close to zero as the fitting sample size grows larger. In nearly all datasets, a mean error less than 0.1 can be reached after the sample size increased to 1,000, suggesting that the proposed adaptive prediction of learning curve is effective. The rate that RMSE or MAE drops is also an indication of data variation and robustness of classifier. Among the three datasets, MAE and RMSE drop the fastest for D3 for all three algorithms. This suggests that D3 has least variation, because it required the least amount of fitting samples to model the behavior of learning curve. Compared among the three algorithms, LDA and GBM are the most robust because their associated RMSE and MAE drop the fastest for all three datasets.

3.2 Calibration Sample Size Prediction

Finally, we investigated the performance of calibration sample size prediction. Recall that our prediction algorithm reports two decisions for samples sizes, namely the predicted baseline sample size $\widehat{S_B}$ and the population-wide upper bound S_β. We first examined $\widehat{S_B}$, where the tolerance level Ts on the 95 % fitting confidence interval was set to 2 %. For each dataset, 1,000 samples were randomly selected as the prior dataset, from which classification AUCs for each of the three classifiers were obtained. Then, the average AUCs of the three classifiers were used as the baseline performance P_B for each dataset. Figure 4 illustrates the trajectories of the adaptive prediction for all cases. Given a baseline performance, we gradually increased calibration data and updated learning curve coefficients. Once we reached a pre-set confidence level for the curve fitting, this algorithm would stop from obtaining new calibration data and provide a prediction of the calibration sample size that will be used in order to reach the baseline. For instance, it was predicted that for dataset D1, at the tolerance level 2 %, $\widehat{S_B} = 880$ samples are needed for GBM to reach the baseline performance $P_B = 0.84$. In this case, the true sample size for $P_B = 0.84$ is 820 and therefore our algorithm predicted 30 more samples or a 7.5 % error. Setting a more stringent tolerance level can further reduce this error. Examining all nine cases, we can always observe convergence to the true calibration samples size as the sample size increases and the prediction result when compared with the baseline performance has an error of 17.4 % on average. More specifically, GBM has an average prediction error of 6.2 %, LDA has an error of 26.5 %, and SVM has an error of 20 %. Besides, it is observed that our algorithm usually overestimated the baseline sample size. In practice, an overestimation is preferred because an overestimated $\widehat{S_B}$ would always ensure that the baseline performance could be reached. As discussed previously, $\widehat{S_B}$ can be used to select classifier for BCI. In this case, LDA is selected for D1 and D2, whereas SVM is selected for D3. These selections are consistent with those based on the true baseline sample size, suggesting again that our prediction algorithm can correctly assess the relationship between sample size and classification performance.

Finally, we investigated the scenarios where we need to consider the population-wide upper bound S_β . Particularly, we used the dataset D1 to simulate three potential scenarios. For the first scenario, we set the baseline AUC $P_B = 0.9$. We used the randomly selected 1,000 samples to estimate the distribution of baseline sample size, from which we had $S_\beta = 270$ (Fig. 5A). Then, we focused on Subject 1 and determined that $\widehat{S_B} = 140$. Since $\widehat{S_B} < S_\beta$, the calibration sample size was predicted to be 140 (Fig. 5A). In the second scenario, we increased the $P_B = 0.95$. Once again, we estimated the distribution of the baseline performance from the 1,000 prior data samples and $\widehat{S_B}$ for Subject 1. This time, we had $\widehat{S_B} = 550$ but predicted calibration sample $\widehat{S_B} = 590$ (Fig. 5B). Since $\widehat{S_B} > S_\beta$, we would suggest to collect 550 samples for calibration. However, notice that $\widehat{S_B}$ is only 40 samples more than S_β, therefore one might consider collecting the predicted 590 samples instead, if resources permit. In the third scenario, we still set $P_B = 0.95$ and therefore $S_\beta = 550$. However, we chose to predict the baseline sample size for Subject 2, where we had $\widehat{S_B} = 1{,}340$ (Fig. 5**C**). Since this time $\widehat{S_B} \gg S_\beta$, we would suggest to collect only 550 samples.

4 Conclusion and Future Work

This study proposed a new algorithm for predicting calibration sample size for EEG)-based classification in BCIs. The key component of the algorithm is an adaptive fitting of a learning curve. Instead of producing a single prediction, our algorithm outputs a predicted baseline sample size and a population-wide upper bound. Empirical results showed that our algorithm can correctly predict the behavior between classification performance and sample size. Providing two predicted sample sizes gives user more flexibility to reach a case-specific decision. In addition, the predicted sample size can be used to select an appropriate classifier for BCI.

Another important future direction is to investigate the integration of the sample size prediction methods with transfer learning to achieve reduced calibration data. There are two potential directions for this investigation. First, we can investigate progressive classifiers with TL and generate a learning curve for sample size prediction based on the results coming from these classifiers. Second, we can exploit TL when a subject with the predicted sample size much greater than our expected baseline calibration sample size. Specifically, we can design TL algorithms for the subject to improve the classification accuracy and also reduce calibration samples.

Acknowledgments. Research was sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implies, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for the Government purposes notwithstanding any copyright notation herein. This work received computational support from Computational System Biology Core at the University of Texas at San Antonio, funded by the National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health.

References

1. Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., et al.: Brain-computer interface technology: a review of the first international meeting. IEEE Trans. Rehabil. Eng. **8**, 164–173 (2000)

2. Bigdely-Shamlo, N., Vankov, A., Ramirez, R.R., Makeig, S.: Brain activity-based image classification from rapid serial visual presentation. IEEE Trans. Neural Syst. Rehabil. Eng. **16**, 432–441 (2008)
3. Wu, D., Lance, B.J., Parsons, T.D.: Collaborative filtering for brain-computer interaction using transfer learning and active class selection. PLoS ONE **8**, e56624 (2013)
4. Sun, S., Zhou, J.: A review of adaptive feature extraction and classification methods for EEG-based brain-computer interfaces. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1746–1753 (2014)
5. Panicker, R.C., Puthusserypady, S., Sun, Y.: Adaptation in P300 brain–computer interfaces: a two-classifier cotraining approach. IEEE Trans. Biomed. Eng. **57**, 2927–2935 (2010)
6. Eng, J.: Sample Size Estimation: How Many Individuals Should Be Studied? Radiology **227**, 309–313 (2003)
7. Suresh, K., Chandrashekara, S.: Sample size estimation and power analysis for clinical research studies. J. Hum. Reprod. Sci. **5**, 7 (2012)
8. Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., Ngo, L.H.: Predicting sample size required for classification performance. BMC Med. Inform. Decis. Mak. **12**, 8 (2012)
9. Zodpey, S.P.: Sample size and power analysis in medical research. Indian J. Dermatol. Venereol. Leprology **70**, 123 (2004)
10. Meek, C., Thiesson, B., Heckerman, D.: The learning-curve sampling method applied to model-based clustering. J. Mach. Learn. Res. **2**, 397–418 (2002)
11. Cortes, C., Jackel, L.D., Solla, S.A., Vapnik, V., Denker, J.S.: Learning curves: asymptotic values and rate of convergence. Adv. Neural Inf. Process. Syst. **6**, 327–334 (1994)
12. Meng, J., Meriño, L.M., Shamlo, N.B., Makeig, S., Robbins, K., Huang, Y.: Characterization and robust classification of EEG signal from image RSVP events with independent time-frequency features. PLoS ONE **7**, e44464 (2012)
13. U.S. Department of the Army. Use of volunteers as subjects of research. AR 70–25 Washington DC. Government Printing Office (1990)
14. U.S Department of Defense Office of the Secretary of Defense, Code of federal regulations, protection of human subjects. 32 CFR 219, vol. 32 CFR 219 (1999)
15. Chuang, S.-W., Ko, L.-W., Lin, Y.-P., Huang, R.-S., Jung, T.-P., Lin, C.-T.: Co-modulatory spectral changes in independent brain processes are correlated with task performance. Neuroimage **62**, 1469–1477 (2012)
16. Sajda, P., Pohlmeyer, E., Wang, J., Parra, L.C., Christoforou, C., Dmochowski, J., et al.: In a blink of an eye and a switch of a transistor: cortically coupled computer vision. Proc. IEEE **98**, 462–478 (2010)
17. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**, 367–378 (2002)
18. McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition, vol. 544. Wiley, New York (2004)