# A Data Analysis Competition to Evaluate Machine Learning Algorithms for use in Brain-Computer Interfaces

Paul Sajda, Adam Gerson, Klaus-Robert Müller, Benjamin Blankertz and Lucas Parra

***Abstract*— We present three datasets that were used to conduct an open competition for evaluating the performance of various machine learning algorithms used in brain-computer interfaces. The datasets were collected for tasks that included 1) detecting explicit left/right (L/R) button press, 2) predicting imagined L/R button press and 3) vertical cursor control. A total of ten entries were submitted to the competition, with winning results reported for two of the three datasets.**

***Index Terms*— Data analysis competition, Brain Computer Interface, machine learning, electroencephalography (EEG)**

## I. Introduction

A variety of machine learning and pattern classification algorithms have been used in the design and development of brain-computer interfaces (BCI). Though many of these algorithms have been reported to give impressive results, it is difficult to assess their relative utility given their evaluation on different data sets and/or using different performance metrics. One approach for comparing various algorithms that has been used by the machine learning community is to conduct data analysis competitions. Such competitions have proven quite successful, for example in assessing algorithms for time-series prediction[1].

In an effort to provide a common, and relevant, set of data for evaluation of algorithms used in BCI we announced a data analysis competition during the Neural Information Processing Systems (NIPS 2001) Brain Computer Interface Workshop (Whistler, Canada, December 2001). Results of this competition were announced at the 2nd International Brain Computer Interface Workshop (Renssellaerville, NY, June 2002). Three electroencephalography (EEG) data sets were provided, each collected for distinct BCI tasks.

Participants in the competition were asked to follow a few simple rules:

1) All data sets should be evaluated single-trial – no averaging across multiple trials.
2) The statistics/metrics outlined in the description of each dataset should be reported.
3) Use of these datasets implies that the participant agrees to cite the origin of the data in any publication.

In the following sections we describe the datasets used in the competition, the classes of algorithms submitted, and the results.[1]

## II. The Datasets

Three datasets were used in the competition. Participants were able to download the data from the web. Each dataset had a set of training trials, with labeled truth data, and a set of test trials. Participants were asked to generate the labels for the test data and submit those to the organizers. The organizers then computed the performance for each participant's entry. Participants were also asked to submit a brief description of the algorithm they used in their analysis.

### A. Dataset 1: EEG self-paced key typing

This dataset was courtesy of Benjamin Blankertz and Klaus-Robert Mueller, Fraunhofer FIRST, and Gabriel Curio, FU-Berlin[2]. This dataset consists of 513 trials of 27 electrode EEG recordings from a single subject. While sitting in a chair, relaxed arms resting on the table, fingers in the standard typing position at the computer keyboard (index fingers at 'f','j' and little fingers at 'a',';') the subject was instructed to press the aforementioned keys with the corresponding fingers in a self-chosen order and timing. The task was to classify EEG potentials as being associated with left or right finger movement. 413 training trials were provided and 100 trials used for testing.

### B. Dataset 2: EEG synchronized imagined movement

This dataset was courtesy of Allen Osman, University of Pennsylvania[3]. The task of each of 9 subjects during the EEG Synchronized Imagined Movement data set was to imagine moving his or her left or right index finger in response to a highly predictable timed visual cue. The goal of competition participants was to classify EEG recordings as belonging to left or right imagined movement. EEG was collected using 59 sensors. 90 trials for each subject (45 labeled left and 45 labeled right) were for training and 90 trials (unlabeled) were for testing.

Paul Sajda and Adam Gerson are with The Department of Biomedical Engineering, Columbia University, New York, NY, USA. E-mail: sajda@columbia.edu, adg71@columbia.edu. Klaus-Robert Müller and Benjamin Blankertz are with Fraunhofer FIRST, Berlin, GERMANY. E-mail: klaus@first.fraunhofer.de, blanker@first.fraunhofer.de. Lucas Parra is with Sarnoff Corporation, Princeton, NJ, USA. E-mail: lparra@sarnoff.com.

[1] More details about the competition, together with the datasets, can be found at http://liinc.bme.columbia.edu/competition.htm.

TABLE I
SUMMARY OF THE THREE DATASETS

| | Dataset 1 | Dataset 2 | Dataset 3 |
|---|---|---|---|
| Task | L/R explicit finger tap | L/R imagined button press | cursor control |
| Collection mode | open-loop | open-loop | closed-loop |
| Classes | 2 | 2 | 4 |
| Subjects | 1 | 9 | 3 |
| Training trials | 413 | 90 | 1152 |
| Testing trials | 100 | 90 | 768 |
| EEG electrodes | 27 | 59 | 64 |

### C. Dataset 3: Closed-loop cursor control

This dataset was courtesy of Gerwin Schalk, Wadsworth Center. The data set consists of 64 electrode EEG recordings from 3 subjects. The task of each subject was to move a cursor on a video screen to 1 of 4 predetermined positions. Each target position differed only in vertical location. Horizontal coordinates were identical for each target position. The objective was to classify EEG recordings as belonging to the correct target position. 1152 trials were available for training and 768 for testing. Note that in contrast to the first two datasets, this dataset was collected closed-loop (i.e. with feedback from the subject). Table I summarizes the three datasets.

## III. Competition Results

A total of ten entries were received. Six entries were received for dataset 1, three entries for dataset 2, and one for dataset 3. These ten entires represent the application of six different algorithms to the three datasets. Table II lists the submitted algorithms and the datasets to which they were applied. Since there was only one submission for dataset 3, we do not report those results.

### A. Results for Dataset 1

Performance for dataset 1 was computed as the fraction correct (fc) classification on the test data. The winning entry was the recurrent neural network by Sottas. Fraction correct on the test set was 0.96. The algorithm began by lowpass filtering the data to 40Hz and then resampling it to 100Hz. A six neuron, fully connected (204 connections) recurrent network, with one readout neuron, was used for classification. The network was setup so that the output neuron responds only after the complete presentation of the input sequence.

Optimization of the connections is performed by a ”dynamic noise annealing” algorithm [4]. This algorithm can be summarized as

1) Add noise (under a given annealing schedule) to the activation dynamics of the internal states of the network (i.e. the 6 neurons).
2) During each learning step, run the network 20 times with noise, giving 20 different possible solutions for the output neuron.
3) An importance function, which is also annealed, is computed and allows allocation of credit or blame to each of these possible solutions.
4) The weights are updated using an EM-type algorithm.

TABLE II
SUMMARY OF ALGORITHM SUBMISSIONS

| Algorithm | Authors | Dataset |
|---|---|---|
| AutoRegressive Model with eXogenous (ARX) Linear Discriminant Analysis | Burke, Kelly, Chazal, and Reilly | 1 |
| Common Spatial Subspace Decomposition (CSSD) and Multiple Electrode Activity Subtraction (MEAS) Linear Discriminant Analysis | Gao | 1, 2, 3 |
| Decision Tree and Neural Network Classifier | Dam, Tosevski, Belista, and El-Ali | 1 |
| Slow Potential Shift nu-Support Vector Classifier with CV Criterion Parameter Tuning | Rosipal, Trejo, and Wheeler | 1, 2 |
| Recurrent Neural Network Optimized by Dynamic Noise Annealing | Sottas | 1 |
| Feature combination using a Fisher discriminant: Combining the Bereitschaftspotential (BP), (adaptive) autoregressive coefficients and Common Spatial Patterns (CSP) | Dornhege, Blankertz, and Zander | 2 |

The performance of this algorithm is comparable with that reported by the contributors of the data [2].

It should be noted that two other entries (Gao and Rosipal et al.) performed very close to the winning entry ($fc = 0.95$). Therefore the difference between the three entries is not likely to be statistically significant.

### B. Results for Dataset 2

Performance for dataset 2 was also computed as the average fc on the test set (averaged across the nine subjects). The winning entry was a feature combination approach using a Fisher discriminant by Dornhege et al. The performance was $fc = 0.76$.

The feature combination approach exploited several motor-related features which are known from the BCI literature:

1) Non-oscillatory Event-Related Potentials (ERPs).
2) Coefficients of (adaptive) autoregressive models (AR).
3) Common Spatial Patterns (CSP).

The classification label, computed by a Fisher discriminant, was estimated for the test data by using the feature combination method that gave the best cross-validation error. These results are slightly better than those reported by the contributors of the data [5].

## IV. Conclusion

We have reported on the BCI data analysis competition announced at the NIPS2001 BCI workshop with results presented at the 2nd International BCI Workshop (2002). Competitive entries were received for two of the three datasets. The third dataset, cursor control, received only a single entry. We believe that this is likely due to the fact that this dataset is collected closed-loop. To evaluate a new machine learning algorithm requires explicitly placing it in loop, since it may potentially change the feedback response from the subject and ultimately the entire system (i.e. machine + human) response. Such a test is not possible with this dataset since it is not possible to “break-the-loop”. Nonetheless, a future challenge for such competitions will be to develop datasets and paradigms which can be used to illustrate generalization of algorithm performance to realistic closed-loop, on-line processing scenarios.

## V. Acknowledgments

This work was funded by the Defense Advanced Research Projects Agency (DARPA).

## References

[1] N. Gershenfeld and A. Weigand, *Time series prediction: forcasting the future and understanding the past*. Addison-Wesley, 1993.

[2] B. Blankertz, G. Curio, and K.-R. Müller, “Classifying single trial EEG: Towards brain computer interfacing,” in *Advances in Neural Information Processing Systems (NIPS 01)*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2002.

[3] A. Osman and A. Robert, “Time-course of cortical activition during overt and imagined movements,” in *Cognitive Neuroscience Annual Meeting*, New York, March 2001.

[4] P. Sottas and W. Gerstner, “Dynamic noise annealing for learning temporal sequences with recurrent neural networks,” in *Artifical Neural Networks-ICANN 2002*, 2002, pp. 144–149.

[5] L. Parra, C. Alvino, A. Tang, B. Pearlmutter, N. Yeung, A. Osman, and P. Sajda, “Linear spatial integration for single trial detection in encephalography,” *NeuroImage*, vol. 15, no. 1, pp. 223–230, 2002.