

THE SIXTH ANNUAL MLSP COMPETITION, 2010

Kenneth E. Hild II¹, Mikko Kurimo², Vince D. Calhoun^{3,4}

¹Dept. of Biomedical Engineering, Oregon Health and Science University, Portland, OR, USA

²Dept. of Information and Computer Science, Aalto University, Finland

³Dept. of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA

⁴The Mind Research Network, Albuquerque, NM, USA

k.hild@ieee.org, mikko.kurimo@tkk.fi, vcalhoun@unm.edu

ABSTRACT

For the Sixth Annual Machine Learning for Signal Processing competition, sponsored by Nokia and PASCAL2, entrants were asked to develop a classifier, with optional feature extraction, that uses electroencephalography (EEG) data collected during an image presentation (visual odd-ball) task and optimally determines whether each image presented to a user contains or does not contain a pre-specified target. In this paper, we (the organizers of the competition) briefly describe the application, the data, the rules, and the outcomes of the competition. A total of 35 teams entered the contest. Training data were provided. The entries were tested using disjoint test data. The three teams with the best performing entries describe the approach they used in three separate companion papers, all of which appear in this year's conference proceedings.

1. INTRODUCTION

This year marks the sixth time the Machine Learning for Signal Processing (MLSP) Technical Committee has hosted a data analysis competition in conjunction with the annual MLSP conference. In previous years the competition consisted of blindly separating instantaneous mixtures of synthetic data and convolutive mixtures of real data (speech), denoising magnetoencephalographic (MEG) data, processing functional magnetic resonance image (fMRI) data, rejecting artifacts in event-related electroencephalographic (EEG) data, maximizing the annual rate of return by trading stocks over a six-month period [1], and applying computer vision methods to find a specific person in a set of images. This year, the goal is to determine, using EEG data, when the user (from whom the EEG data are collected) detects an instance of a pre-specified target while they view a series of images presented on the screen in front of them.

Brain machine interfaces (BMI's), also called brain computer interfaces (BCI's), infer the intentions of a user based on neural signals. The goal of BMI research is either

to restore the ability of users having functional deficits (i.e., those who have physical impairments that limit their ability to communicate, ambulate, or control various devices, such as a computer mouse) or to augment the ability of users having no such deficits. BMI's are being developed at institutions all over the world [2]-[6]. Another indication of the widespread interest in BMI's is found in the list of entrants, which is given below. In the next section, we briefly describe the training data and the application. Afterwards, we describe the test data and the competition rules. Finally, we list the entrants and show how well each entry performed.

2. TRAINING DATA AND APPLICATION

The training data, as of the date this paper was submitted, may be downloaded from <http://www.bme.ogi.edu/~hildk/mlsp2010TrainingData.mat> (44 MB). The training data, which is in Matlab[®] format, consist of EEG data collected while a subject viewed satellite images that were displayed in the center of an LCD monitor approximately 43 cm in front of them. There are 64 channels of EEG data. The total number of samples is 176378. The sampling rate is 256 Hz. There are 75 blocks and 2775 total satellite images. Each block contains a total of 37 satellite images, each of which measures 500 x 500 pixels. All images within a block were displayed for 100 ms and each image was displayed as soon as the preceding image was finished. Each block was initiated by the subject after a rest period, the length of which was not specified in advance. The subject was instructed to fixate on the center of the images and to press the space bar whenever they detected an instance of the pre-specified target. Subjects also needed to press the space bar to initiate a new block and to clear feedback information that was displayed to the subject after each block.

We expect a particular neural signature, the P300, to occur whenever the subject detects an instance of a target in one of the satellite images. The P300 gets its name from the fact that detection of a (rare, pre-defined) target causes

Table 1. Variables in the training data.

Variable	Definition
eegData	The 64-channel EEG data.
imageTrigger	Values of 1 correspond to the onset of non-target images. Values of 2 correspond to the onset of target images. Values of 0 are used elsewhere.
buttonTrigger	Values of 1 correspond to the onset of a button press.
eegLabel	The label for each of the 64 electrodes.
eegCoord	The coordinates of each of the 64 electrodes. The (spherical) coordinates are measured in degrees of inclination from Cz (positive values correspond to the right hemisphere, negative values correspond to the left hemisphere) and degrees of azimuth (from T7 for the left hemisphere and from T8 for the right hemisphere, positive values correspond to anti-clockwise rotations, negative values correspond to clockwise rotations).

a deflection of the EEG signal approximately 300 ms after the visual stimulus is presented [7]. The P300 is particularly prominent in the midline channels (e.g., Pz, Cz, Fz; see Figure 1). In addition, there is a separate neural signature associated with the pressing of the space bar [8],[9]. For our data, this second neural signal occurs around 500-600 ms after the stimulus containing the target is presented. The variables in the training data are defined in Table 1.

3. TEST DATA AND COMPETITION RULES

The test data are similar to the training data, but there are some important differences. As before, there are 64 channels of EEG test data, the sampling rate is 256 Hz, there is no delay between images within the same block, the subject rests between blocks for as long as they wish, and the subject presses the space bar to signify when they detect a target, to initiate a new block, and to clear feedback information that was displayed after each block.

Unlike before, the test data consist of 890 blocks and 9891 satellite images, the total number of samples of EEG data is 1603334, every other image within a block is a mask image (mask images do not contain targets), the buttonTrigger variable is not available, and the imageTrigger variable takes only values of 0 or 1, where a 1 corresponds to the onset of each prospective target image (i.e., the satellite images) and 0 is used elsewhere. Another difference is that 4 different image durations are used in the test data. The image durations, which apply to both satellite and mask im-

ages, are (approximately) 50 ms, 100 ms, 150 ms, and 200 ms. All images within a given block have the same image duration and all blocks having a specified image duration are grouped together. Each block contains 22, 10, 7, and 5 prospective target images when the image duration is 50 ms, 100 ms, 150 ms, and 200 ms, respectively. Keep in mind that the time difference between successive prospective target images is twice the corresponding image duration due to the presence of the mask images. Hence, successive prospective target images within a block appear every (approximately) 100 ms, 200 ms, 300 ms, or 400 ms.

The test data were not available to the participants. Instead, participants submitted Matlab[®] code, which the competition chairs tested by passing the test data to the submitted code. Successful submissions consisted of: (1) the names of the team members (each person may belong to at most two teams and each team is allowed a single submission), (2) the name(s) of the host institutions of the researchers, (3) a 1-3 paragraph description of the approach used, and (4) Matlab[®] code, myFunction.m, which is called using “[out] = myFunction(eegData,t,imageTrigger,eegLabel,eegCoord).” Entrants were also allowed to supply one data file from which parameter values could be read and they were instructed that the submitted code must not write to any drive, must finish running in a reasonable time, and must consist only of regular, uncompiled Matlab[®] code (P-code, mex files, and compiled code are, e.g., forbidden). The deadline for submitting to the competition was April 8, 2010.

The 2010 MLSP Competition Committee is pleased to announce that we will award prizes to the winners at the conference in Kittila, Finland (awards are given most years, but they are not always advertised ahead of time). This year, the planned awards consist of three N900 high-performance mobile computers, courtesy of Nokia, and a total of 1500 Euros to be distributed as travel grants, courtesy of the PASCAL2 Challenge Program. The selection of the winners is based on: (1) the AUC performance of the submitted methods, which is described below, and (2) the requirement that at least one member of each selected team attend the 2010 MLSP Conference. In addition, members of the 2010 MLSP Competition Committee (and everyone belonging to any of the labs of the 2010 MLSP Competition Committee) are not eligible for these awards.

4. RESULTS

The output of each submitted algorithm (i.e., the probability that each satellite image contains a target) was used to rank the likelihood that each exemplar contained a target. Using the ranked list, we constructed a receiver operating characteristic (ROC) curve, from which we computed the AUC performance metric (the area under the ROC curve).

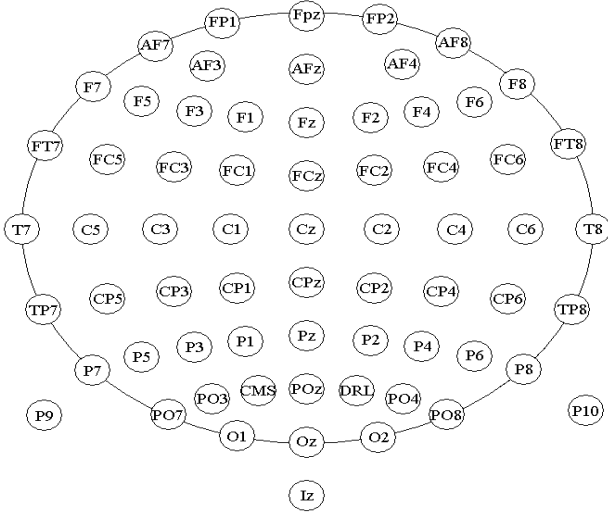


Fig. 1. Top view of the locations of the 64 electrodes.

For this year’s competition a total of 35 algorithms were submitted. There were many different types of approaches used by the entrants. It is generally difficult to categorize the methods according to the field of research in which it is most commonly used. However, we did notice that several of the methods used tools that are commonly used for face recognition or speech recognition. Of the 35 entries, 11 used some form of support vector machine (SVM) classifier for the final classification, 9 used linear discriminant analysis (LDA; also known as Fisher discriminant analysis), 5 used a hidden-layer neural network, and 3 used a linear logistic classifier. Hence, the remaining 7 methods used a classifier other than the ones listed above. We report in Table 2 the mean performance, averaged over entries, for each of these groups of classifiers. We include these averages for informational purposes only. Because proper controls were not instituted (the algorithms differ in many ways besides the classifier), the relative classifier performance given in Table 2 is not expected to generalize to other datasets.

Bootstrap aggregating, which is also known as bagging, is a mechanism by which the outputs of multiple, possibly different, classifiers can be combined into a single classification function [10]. It is possible to use this meta-analysis technique with any classifier. A total of 9 entries used either bagging or a technique that is very similar to bagging. Of the 9 entries that used bagging: 4 used an SVM as the base classifier, 3 used LDA, 1 used linear logistic, and 1 used a radial basis function (RBF). Table 2 includes the mean AUC performance of the group of entries that used bagging and those that did not. No clear patterns emerged to indicate which classifier or pre-processing method is preferred in general. However, we can make a few specific

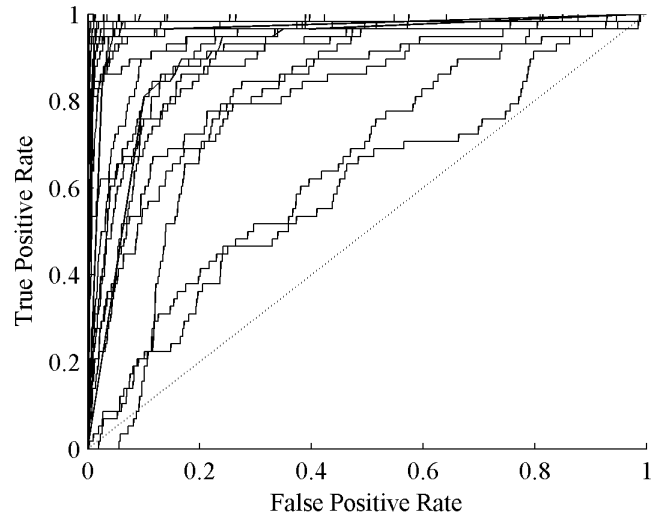


Fig. 2. ROC curves (one for each team) for the training data.

Table 2. Mean and standard deviation of classifier performance for each type of classifier.

Classifier	No. of entries	Mean AUC	Std
SVM	11	0.64	0.12
LDA	9	0.70	0.13
Neural network	5	0.66	0.11
Linear logistic	3	0.63	0.24
Other	7	0.67	0.15
Bagging	9	0.62	0.14
Non-bagging	26	0.68	0.13
All classifiers	35	0.66	0.13

comparisons. Five of the methods were very similar except for the classifier. For these 5, the SVM classifier (with Gaussian kernel) performed slightly better than LDA, linear SVM, and a convolutional neural network, and all 4 of these performed much better than a hidden-layer neural network. Likewise, 2 of the methods were very similar except for the temporal window. The method using a window of 0 to 500 ms performed slightly better than the method using a window of 0 to 1000 ms. We also noticed that the performances of arguably the least and most complex methods submitted were below the performance corresponding to chance.

The classifiers used in the 3 best entries are: a generative classifier (which estimates a joint probability function) [11] with an AUC of 0.8229, a classifier based on T-weights [12] with an AUC of 0.8217, and an SVM with a Gaussian kernel [13] with an AUC of 0.8188. Each of the top 3 entries uses band-pass filtering. The mean lower-cutoff frequency of the top 3 entries is 0.8 Hz. The mean upper-cutoff frequency of

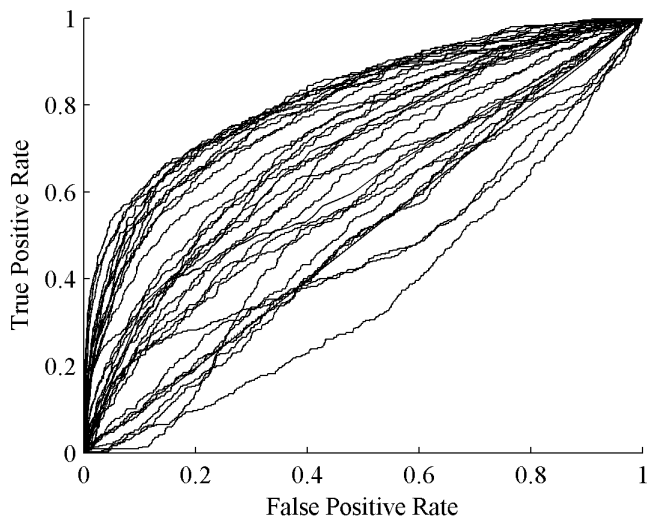


Fig. 3. ROC curves (one for each team) for the test data.

Table 3. Members of the top three teams.

#	Name(s)	Affiliation(s)
1	Jose M. Leiva Suzanne M.M. Martens	Univ. Carlos III de Madrid Max Plank Inst. Biol. Cyb.
2	Zafer Iscan	Istanbul Technical Univ.
3	Benjamin Labbe Xilan Tian Alain Rakotomamonjy	INSA de Rouen INSA de Rouen Universite de Rouen

the top 3 entries is 12 Hz.

Figure 2 shows the ROC curves (one for each team) for the training data, Figure 3 shows the ROC curves for the test data, and Figure 4 shows the performance metric obtained by each team (sorted from largest to smallest). Keep in mind that it is entirely possible that the submitted code did not perform as expected for trivial reasons relating to a misunderstanding between what was submitted and what was expected (however, we did check to make sure that all methods achieved an AUC of at least 0.60 on the training data). Notice that there is little difference between the performances of the top several teams. Table 3 lists the names of the members, and their corresponding affiliation(s), for each of the top three teams. A summary of the method used by each of the top 3 teams can be found in the conference proceedings [11]-[13]. Table 4 lists the affiliation(s) of each of the 35 teams, their rank, and their AUC performance on the test data. In this table, the values for AUC performance are reported after rounding to the nearest hundredths.

Table 4. Rank, affiliation(s), and AUC for each team.

#	Affiliation(s)	AUC
1	Universidad Carlos III de Madrid, Max Plank Institute for Biological Cybernetics	0.82
2	Istanbul Technical University	0.82
3	INSA de Rouen, Universite de Rouen	0.82
4	Stanford University	0.81
5	Universite de Rouen, INSA de Rouen, Universite de Montreal	0.81
6	Universite de Rouen	0.81
7	Universita degli Studi di Milano	0.80
8	Universite de Rouen, INSA de Rouen	0.80
9	Donders Inst. Brain, Cognition and Behaviour	0.79
10	University of Greifswald	0.79
11	Bielefeld University, Honda Research Institute	0.78
12	Brain Science Institute, Riken	0.77
13	Oregon Health and Science University	0.75
14	Nara Institute of Science and Technology	0.72
15	University of Southern California	0.71
16	University of Southern California	0.70
17	Univ. Desarrollo Tecn., Innovacion en Comun., Universidad de Las Palmas de Gran Canaria	0.69
18	University of Ulm	0.68
19	University of Reading	0.68
20	Univ. Autonoma Metropolitana Iztapalapa, Universidad Nacional de Entre Rios	0.66
21	Fraunhofer Institute FIRST	0.64
22	INSA de Rouen, Universite de Rouen	0.64
23	University of Essex	0.62
24	Nara Institute of Science and Technology	0.62
25	Universite Catholique de Louvain	0.59
26	Yonsei University	0.56
27	State University of New York at Buffalo	0.53
28	Katholieke Universiteit Leuven	0.53
29	Advanced Brain Signal Processing, Riken	0.50
30	Yahoo Research Barcelona	0.50
31	Grenoble University	0.50
32	Ghent University	0.49
33	University Federico II of Naples	0.48
34	Carnegie Mellon University	0.46
35	University Carlos III Madrid	0.37

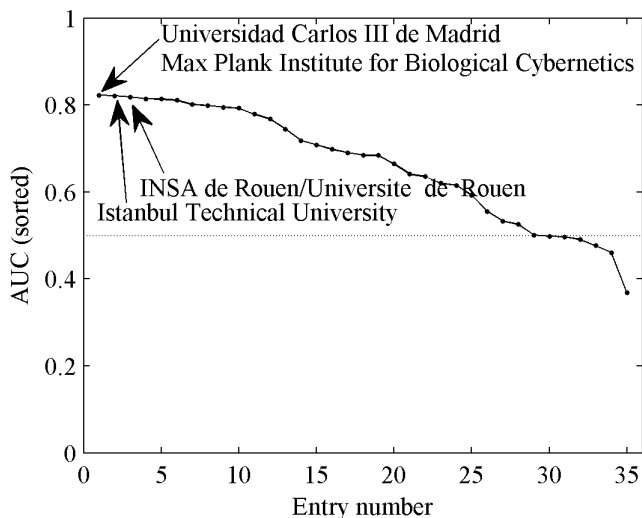


Fig. 4. AUC for each team (sorted from largest to smallest).

5. CONCLUDING REMARKS

Users are easily able to detect objects shown at image durations of 50 or 100 ms when the targets are sufficiently distinguishable from the distractors. For the images used in the test, however, some of the targets are difficult to detect even if the image duration is increased to 200 ms. The user (from whom the test data were collected) detected 74% of the targets. For comparison, for a false positive rate of 0.10, the top performing algorithm detected 60% of the targets (keep in mind that, for small false positive rates, the performance of the machine learning algorithm is essentially limited to the detection rate of the user). The highest AUC reported here, 0.82, matches the performance obtained by the researcher who originally collected the data. This level of performance has proven somewhat useful in the particular application for which the data were collected.

We had a very good response to the competition this year. As can be seen from Table 4, the host institutions of the entrants are located in countries including (in no particular order): Spain, Germany, Turkey, France, United States, Italy, Canada, Japan, United Kingdom, Netherlands, Belgium, South Korea, Mexico, and Argentina. We would like to once again thank the entrants for submitting their algorithms. We have enjoyed organizing the competition this year and we are looking forward to next year's competition.

6. REFERENCES

[1] K.E. Hild II, V. Calhoun, "The Fourth Annual 2008 MLSP Competition," Intl. Conf. on Machine Learning

for Signal Processing (MLSP '08), Cancun, Mexico, pp. 38-42, Oct. 2008.

- [2] J.P. Donoghue, L. Hochberg, "Designing a Neural Interface System to Restore Mobility," in *Neuromodulation*, Elliot S. Krames, P. Hunter Peckham, and Ali R. Rezai, Eds., Academic Press, Burlington, MA, pp. 229-242, 2009.
- [3] W.L. Woon and A. Cichocki, "Novel Features for Brain Computer Interfaces," *Journal of Computational Intelligence and Neuroscience*, pp. 1-7, June 2007.
- [4] J.C. Sanchez, J. Principe, *Brain-Machine Interface Engineering (Synthesis Lectures on Biomedical Engineering)*, Morgan & Claypool Publishers, San Rafael, CA, Nov. 2006.
- [5] C. Sannelli, T. Dickhaus, S. Halder, E.M. Hammer, K.R. Muller, B. Blankertz, "On optimal channel configurations for SMR-based brain-computer interfaces," *Brain Topography*, Feb. 2010.
- [6] A. Yazdani, J.-S. Lee, T. Ebrahimi, "Implicit Emotional Tagging of Multimedia Using EEG Signals and Brain Computer Interface," *Proc. of ACM Multimedia*, Workshop on Social Media, Beijing, China, Oct. 2009.
- [7] T.W. Picton, "The P300 wave of the human event-related potential," *Journal of Clinical Neurophysiology*, Vol. 9, No. 4, pp. 456-479, 1992.
- [8] K.B. Campbell, E. Courchesne, T.W. Picton, K.C. Squires, "Evoked potential correlations of human information processing," *Biological Psychology*, Vol. 8, pp. 45-68, 1979.
- [9] K.H. Chiappa, *Evoked Potentials in Clinical Medicine*, 3rd ed., Lippincott-Raven, Philadelphia, PA, 1997.
- [10] L. Breiman, "Bagging predictors," *Machine Learning*, Vol. 24, No. 2, pp. 123-140, 1996.
- [11] J.M. Leiva, S.M.M. Martens, "MLSP Competition, 2010: Description of first place method," *Machine Learning for Signal Proc. (MLSP '10)*, Kittila, Finland, Sept. 2010.
- [12] Z. Iscan, "MLSP Competition, 2010: Description of second place method," *Machine Learning for Signal Proc. (MLSP '10)*, Kittila, Finland, Sept. 2010.
- [13] B. Labbe, X. Tian, A. Rakotomamonjy, "MLSP Competition, 2010: Description of third place method," *Machine Learning for Signal Proc. (MLSP '10)*, Kittila, Finland, Sept. 2010.