“Machine Learning and the μ Rhythm”

Akram Helou
Dave Thompson

Introduction

Brain-Computer Interfaces (BCIs) offer a possibility of communication to individuals with severe paralysis or advanced neurodegenerative diseases. One common form of non-invasive BCI is based on Sensori-Motor Rhythms (SMR) known as the μ and β rhythms. These rhythms are “resting rhythms” observed over motor cortex, and are interrupted by real or imagined movement. The μ rhythm’s frequency content is mostly in the 8-12 Hz band, while the β rhythm is closer to 18-26 Hz [1]. With training, a human can increase his or her control of the μ rhythm, and it can be used to control cursor movement in several dimensions.

The μ and β rhythms are typically measured by surface electro-encephalogram (EEG). EEG is safe, inexpensive, and portable, but has a very poor signal-to-noise ratio relative to many other BCI technologies. Because user training to produce reliable μ rhythm signals is time-consuming and the outcome is uncertain [2], adaptive signal processing and machine learning can offer substantial benefits [3].

Background/Related Work

Only a few BCI laboratories actually run subjects – instead, most concentrate on signal processing approaches. While machine learning techniques have been applied in a variety of settings with varying effect [3-8], one of the pioneering labs of the μ-rhythm still recommends simple linear techniques (publishing [1] as late as 2005). Although the Berlin BCI lab does combine subject work with advanced machine learning [3], Wadsworth’s tutorials (written to get a new BCI lab up to speed) still advocate picking single features based on r^2 values [9].

Support Vector Machines (SVMs), in particular, have been applied with varying results. In the BCI competitions, SVMs have done very well (such as in BCI competition III, data set V), but have also performed very poorly (such as when, in data set IVa of the same competition, the last four places were all SVMs and the best SVM-based classifier was 25% behind the winner) [4]. As the organizers of the competition recognize, there is a great deal of variance in how much effort and expertise is applied in different submissions, and the results cannot be considered a fair comparison of methods [4]. However, the often-poor performance of advanced machine learning techniques in competition could be one reason why their successes are overlooked. Also, the number of subjects and sessions in the BCI competitions is limited, so there is no guarantee that the findings will generalize. Finally, one of the organizers has commented in person that the online results, based on simple linear techniques, typically would have won or placed second in the competitions – while requiring far less computational resources.

Common Spatial Patterns (CSP) analysis is a commonly-used machine learning technique for μ-rhythm studies [5]. Unfortunately, to limit the scope of this study, CSP analysis was not performed. Prior to publication of these results, CSP should likely be added for comparison.

This present work will focus on SVM- and Dynamic Time Warping- (DTW) based classification as alternatives to the University of Michigan’s Direct Brain Interface (UM-DBI) project’s current methodology. For comparison, results based on linear regression as in [1] will also be included. The dataset consists of 6 subjects. The experimental design and dataset are described under Methods. Positive results on this independent dataset would lend additional credence to machine learning in μ-rhythm BCIs.

One of our initial observations is that most methods used in a setting where data was collected from many subjects across multiple days were rather simple such as LS regression [1] or simple weighting of the μ and β rhythm based on visual inspection of both signals [10]. At the same time, more sophisticated methods such as SVM were employed in the BCI competitions and led to good results. However, the datasets in the BCI competition are very limited. Therefore, our first aim was to investigate whether sophisticated learners can do equally well if not better than simpler methods on an independent, sizable dataset.

Another important observation concerning previous work is that training and testing mostly occurred on the same day (session) for a single subject ([1] is a notable exception, but is a simple technique). Thus, the question of whether the model trained on one or a few temporally close sessions can generalize across future sessions has not been satisfactorily addressed.

Methods

The following three sections will explain the experimental setup details of our SVM and DTW implementations. Important terms are presented in italics upon first use.

Experimental Design and Data Organization

The dataset comes from the UM-DBI µ training protocol. In this protocol, subjects came in on twenty individual occasions (*sessions*), during which they performed nine 3-minute *runs* of a ball dropping task. In between runs they were allowed short breaks. Each run consisted of 27 *trials* or "ball drops." An example of a trial is shown in **Error! Reference source not found.**. At the start of the trial, a target appeared on one half of the bottom edge of the screen. The ball started at the top center and dropped at a constant vertical speed. Online estimates of µ activity controlled the horizontal movement. Users were asked to imagine right-hand movement (lowering µ energy) for right targets, and relax for left targets.

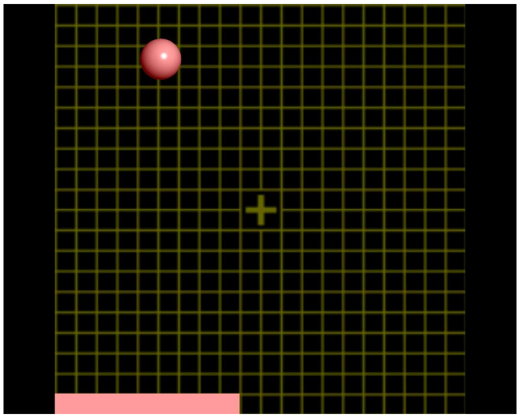

Figure 1 - The test environment of the UM-DBI µ training protocol

EEG samples were taken at 256 Hz by a 16-channel EEG amplifier. Every 8 samples (a *block*), power spectral density (PSD) estimates were calculated according to the maximum entropy method for spectral estimation. The PSD estimates were calculated as 3-Hz bins from 0 to 30 Hz. The total available feature set then included 11 frequency components for each of the 16 channels, or 176 features. Because the trials were approximately 3 s long, each trial contains 91 blocks. This motivates two separate measures of accuracy: *block accuracy* is the percentage of blocks correctly classified by trial target, while *task* or *trial accuracy* is the percentage of trials correctly classified based the results of all blocks.

It is important to note that while the feature set is large, the number of usable features is much smaller. Activity in the µ and β bands has defined spatial locations and known frequency components. Channel 5 is in the ideal spatial location, though channel 9 may be close on some subjects due individual variations. The frequency bands around 9, 12, 18, 21, and 24 Hz were commonly used in this study, as they encompass the known frequencies for µ and β rhythms.

SVM

We explored using SVM under multiple settings for a limited subset of the dataset given. In the end we only employed a few of these settings on the entire dataset, due to the computational requirements of SVMs. We have experimented with different methods for model selection as well as different features based on the EEG PSD estimates for the μ and β rhythm.

We used LIBSVM [11] as our implementation for SVM. For model parameters selection we used either 2 or 5 fold cross validation where the validation sets were randomly chosen. When using a RBF kernel, we used this cross validation to find the best C (cost parameter of the error term) and gamma (bandwidth) parameters. We investigated both RBF and linear kernels. For the feature set, we used the μ and β rhythm PSD estimates on channels 5 and 9 for a total of 8 to 12 features depending on how a wide of a frequency band we considered. In addition to these features, we added a time feature which indicated the elapsed time in blocks since the start of the trial. The reasoning behind this is that without the time feature we are assuming that the PSD estimates are similar at all points during the drop. However, an inspection of the distributions of these signals shows that they vary within a trial. Finally, we experimented with transforming the original μ and β rhythms features into "history features" that take into account the subject's previous control intention. This new set of features was motivated by the inherently imperfect control that even subjects have over their μ and β rhythms.

The long term aim of this work is to methodically train a control model for any subject. That is why we tried to have a uniform training methodology for all subjects. Our reasonable assumption, backed from previous work [1] [10], that subject control over μ and β rhythms occurs over an extended period of time, and thus across sessions, motivated our training methodology. Specifically, we train on blocks from session *i* and then test on the following sessions (Experiment 1). About one fourth of the sessions are reserved for final testing of our model. Model parameter selection using cross validation is performed before final training on the blocks from session *i*. This training methodology allows us to determine at what point in time we think our subject has achieved satisfactory μ and/or β rhythm control and whether there is generalization from one session to subsequent sessions. Another training methodology that we used is a simple variation where instead of training on just session *i*, we also train on all previous sessions (Experiment 2).

DTW

Dynamic Time Warping (DTW) is a technique for measuring the dissimilarity of two waveforms. It allows shifts and scaling in the time domain to be penalized very lightly, while amplitude or shape variations are penalized heavily. In some sense, DTW is a "featureless" approach because only pair-wise distances are calculated and an individual pattern has no feature value associated with it. DTW was applied in both the time and frequency domain. DTW was used in the time domain on the assumption that trials containing μ rhythm have a characteristic shape (which may be shifted or scaled slightly in time). In the frequency domain, the motivation was that the peak frequency might change slightly, but the characteristic shape of the μ rhythm frequency distribution should be present in relaxation trials.

DTW is not, however, a classification method. Ideally, some form of clustering could be used to establish several signal archetypes. Classification could then be performed by matching incoming patterns to the closest signal archetype and adopting that label. Unfortunately, the choice of clustering method (and associated parameters) is a research problem that could not be addressed fully during this study.

For time-based DTW, a simple k-nearest-neighbor classifier scheme was used because the number of training points is small (each trial is a single training point). While performance might be improved by clustering (and certainly computational burden would be decreased), we submit that the k-nearest-neighbor classifier performance can give an intuition of how well DTW-based methods will work.

With frequency-based DTW, nearest neighbor classification failed completely, so we investigated the clustering approach. We implemented hierarchical clustering with complete linkage. A total of 10 clusters was chosen by inspection of the dendrogram of one subject. The prototype from each cluster was a randomly selected member, and the label was the most common class in the cluster.

Results - SVM

We start with the main results of two experiments. Both these experiments used an RBF kernel and relied on two-fold cross validation to find the best C and gamma parameters. Additionally, both experiments used data from six subjects. Each validation fold consisted of randomly selected blocks. In the first experiment, training occurred on all blocks in session *i* and testing was done on subsequent sessions except those designated as part of the final testing set. For each training iteration, cross validation was conducted on session *i* as previously described. In the second experiment, training occurred on all blocks in sessions 1 to *i* and testing was done on subsequent sessions except those designated as part of the final testing set. Cross validation was only conducted on the block samples from the first four sessions. The standard features were used as input to the SVM. These included a total of eight PSD estimates for the μ and β rhythms on channels 5 and 9. In the interest of space and since there are two broad classes of subjects with an equal number of subjects in each, we will focus our attention on only two exemplary subjects and then attach the results of all subjects in a separate document.

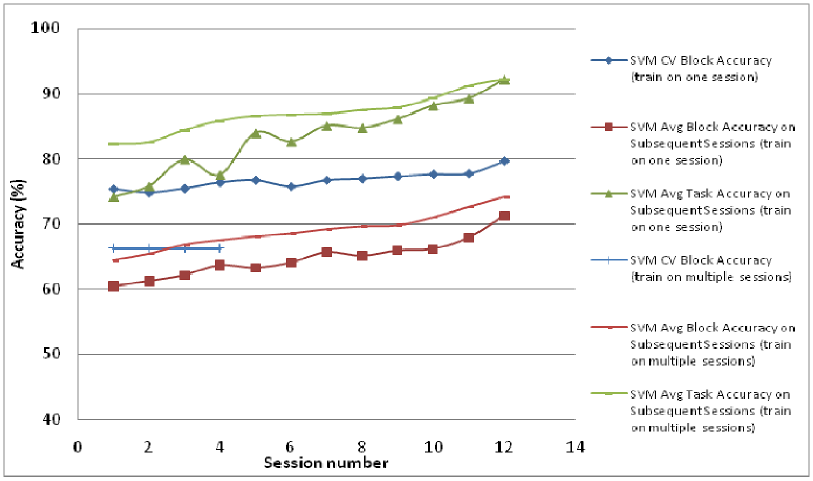

Figure 2: SVM results for Subject 103

Figure 2 shows SVM’s performance on subject 103 under the first and second experiments. Focusing on experiment one’s setting, we notice that while we obtain high CV accuracies for training on a single session, these accuracies do not generalize to subsequent sessions, as evidenced by the red line with squares. This indicates that while a single session’s PSD estimates are somewhat homogenous, this is not the case for PSD estimates across different sessions. This may be because levels of energy vary from day to day based on arousal, alertness, cap position and other factors. This observation led us to consider the second training methodology used in the second experiment, where we hoped to capture all physiological variability. Indeed, in the second experiment we notice the CV block accuracies obtained from training on blocks sampled from the first four sessions are much closer to the block accuracies obtained from testing on subsequent sessions. In addition, the SVM model generalized better to following sessions as evidenced by the on-average 4% jump in block accuracy over results from experiment 1. Additionally, we obtain an improvement of up to 10% on task accuracy. Looking at the lines representing the task accuracies for experiment 1 and 2 (green line with a triangle and green line without a triangle, respectively), it appears that both accuracies are converging. However, this observation is misleading. We should recall that as we move along the x-axis and the session numbers increase then there are less subsequent sessions to test on. Not only this, but the subject is improving control which means that the PSD estimates are becoming more discriminatory and thus future sessions are becoming more and more homogeneous. Overall, under the training setting of experiment 2 we achieve a task accuracy that is increasingly better as we add more sessions, and ranges from 82% to 92%.

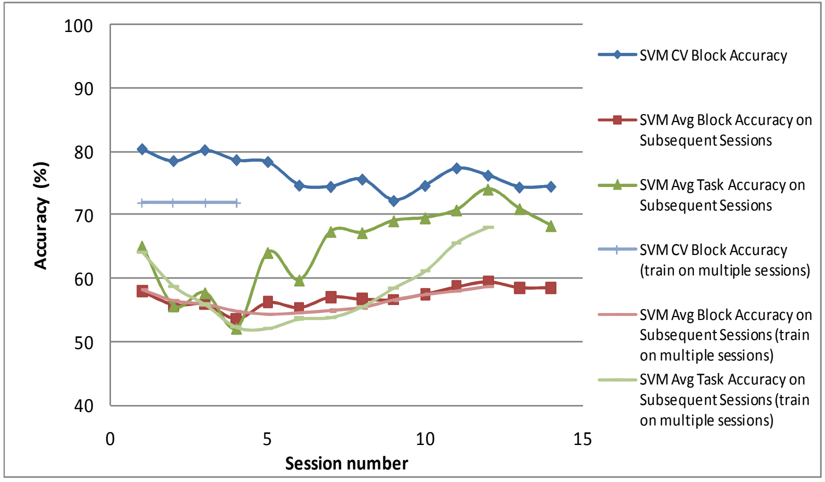

Figure 3: SVM results for subject 126

Figure 3 shows SVM's performance on subject 126 under the first and second experiments as described above. Focusing on experiment one's setting, we again notice that while we obtain high CV accuracies for training on a single session, these accuracies do not generalize to subsequent accuracies as evidenced by the red line with squares. What is initially surprising is that the training methodology of the second experiment does not lead to a significant improvement in generalization. Indeed, the CV accuracies obtained from training on the blocks sampled from the first four sessions overestimate the accuracies obtained from training on subsequent sessions by more than 10%. In addition, the model from experiment two generalized to the following sessions about as well or worse than the model from experiment one, as evidenced by the overlap of the red line with squares and the pink line. We observe a deterioration of up to 15% on task accuracy compared to experiment 1 for the first 7 sessions. This may seem alarming at first because we have been implicitly promoting the superiority of the training methodology used in the second experiment. However, this significant deterioration in performance can be explained by interpreting the SVM's average task accuracy on subsequent sessions when just training on one session as shown by the green line with triangles. We see that for the first four sessions, or the same sessions we did CV on, the model obtained from session *i* is simply not generalizing forward. We propose two possible causes. First, subject 126 still has not achieved satisfactory control over her μ and β rhythm. Second, subject 126 is doing the wrong task, which is difficult to verify given the physical and physiological inactivity of operating a BCI. Further proof that this is the right explanation can be gleaned from the sudden increase (light green line) of the task accuracy when training on single sessions right after session 4. Overall, under the training methodology of experiment 1, we achieve task accuracies that range from 52 (slightly better than chance) to 74%.

Our SVM results from subject I126 motivate a third training methodology which would combine the strength of the first and second methods. We propose to first train an SVM using the 1st methodology. This would permit investigators to indirectly identify sessions where the subject has not achieved good control over her μ and β rhythm or/and is doing the wrong task. We then apply the second methodology as before, except that we exclude training on those sessions that were identified as questionable in the first step. Unfortunately, due to the limited timeframe of this project we could not implement and test this training method.

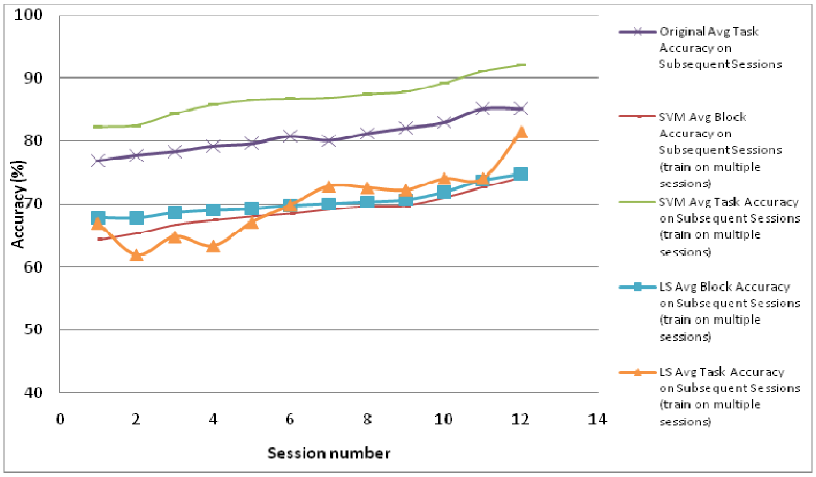

Figure 4: Comparison of SVM results to LS regression and original classifier for subject 103

Figure 4 compares the results obtained from training a model using SVM (experiment 2), LS regression (training method analogous to experiment 2), and the original classifier for subject 103. The task accuracy obtained from SVM is on average 6% better than the one obtained when applying the original classifier and 16% better than the one obtained when applying LS regression. Surprisingly, SVM manages to obtain higher task accuracy than LS regression while having comparable block accuracy. We cannot determine exactly why this is the case. However, it is very likely that LS regression is doing exceedingly well on the immediately following sessions but not as well on sessions far into the future. To obtain high task accuracy it is important to maintain high block accuracy on all subsequent sessions.

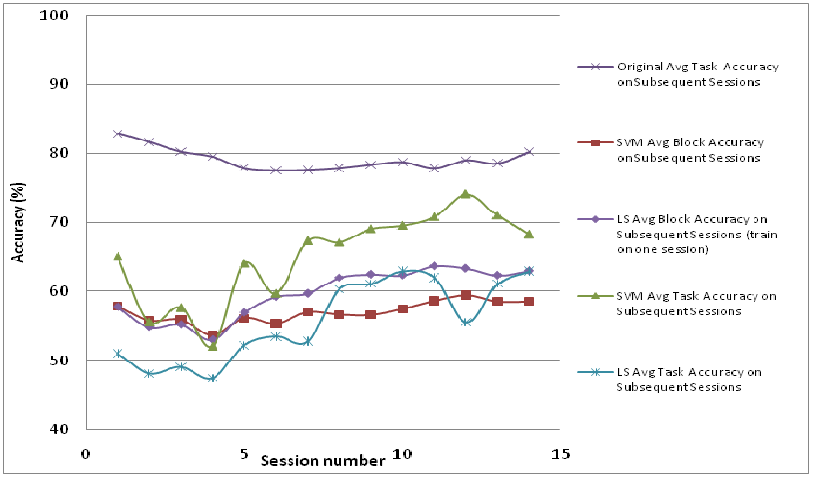

Figure 5: Comparison of SVM results to LS regression and original classifier for subject 126

Figure 5 compares the results obtained from training a model using SVM (experiment 1), LS regression (training method analogous to experiment 1), and the original classifier for subject 126. We see that the mediocre performance of SVM on subject 126 previously discussed shows when compared to the original classifier used in UM-DBI. The task accuracy obtained from SVM is on average 10% higher than the LS regression, but 14% worse than the one obtained when applying the original UM-DBI classifier. Again, SVM manages to obtain higher task accuracy than LS regression while having comparable block accuracy. It is surprising that the results of subject 126 show that a simple weighting of the μ and β rhythms features without training beats two learning algorithms. It is possible that the labels for subject 126 are vastly incorrect, that the subject did not achieve good control, or that the subject was doing the wrong task. Notice how the task accuracy of the original classifier (violet line with Xs) is decreasing instead of increasing as I103 did.

Results – DTW

Unfortunately, the DTW results were not ideal. Results from two subjects are shown in Figure 6, while the rest are included in the appendix. While the DTW-based methods occasionally surpass the original trial accuracy, the majority of the time the original method wins by a substantial margin. Note that the accuracies on Figure 6 are more comparable to the cross-validation accuracies in the previous figures, because each “classifier” was trained on the first six runs on a session and tested on the last three. No between-session comparison was performed.

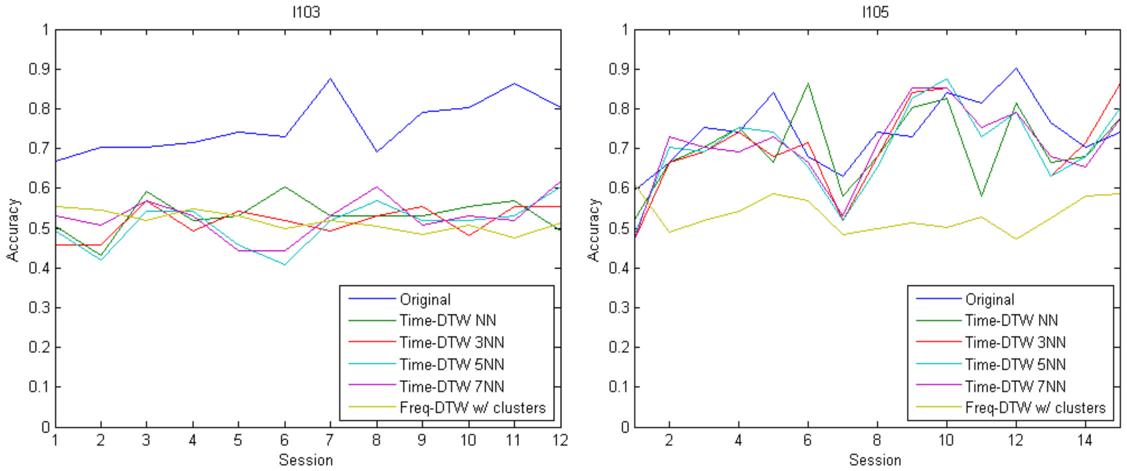

Figure 6 - Results from DTW-based classification on two subjects. (Left) A subject for whom DTW-based methods are not suggested. (Right) A subject who shows promise for DTW-based classification.

Conclusions

SVMs appear to have done fairly well on the dataset. It is interesting to note that regression as in [1] was often out-performed by the original method based on prior knowledge. While it is possible there was an error in the code, this may be because [1] used only experienced, trained subjects whose μ-rhythm control was very stable. They mention in their discussion section that the method may not work well on naïve subjects, and we present here evidence that it does not. LS regression certainly has no inherent feature selection.

The performance of DTW was disappointing. See the following section for ideas on what could have contributed to the poor performance of this method.

Limitations/Future Work

Both investigations have many unanswered questions. On the SVM side, training on all previous sessions may not have been ideal. The first session or two of data include more errors than later sessions, and perhaps excluding them would help performance. Also, our additional features have yet to be tested fully.

On the DTW side, the investigation was extremely limited. Only within-session results were considered, with no attempt to investigate between-session transfer. This was primarily due to computational (and storage) limitations, but without between-session testing it is impossible to prove the practicality of the results. Also, the clustering methodology deserves a research project of its own – determining what linkage to use, how many clusters to allow, and how best to form a cluster archetype, are all open questions.

Finally, the DTW investigation would have benefited from a richer feature set in the frequency domain. There were only 11 frequency points in the PSDs, making the justification for applying DTW in the first place (slight shifts in peak frequency) rather shaky. In future work, DTW could be applied in the frequency domain after re-computing the spectral estimates with finer resolution; we suspect this would produce much better results.

REFERENCES

[1] D. J. McFarland and J. R. Wolpaw, "Sensorimotor rhythm-based brain-computer interface (BCI): feature selection by regression improves performance," *IEEE Trans. Neural Syst. Rehabil. Eng.,* vol. 13, pp. 372-379, Sep. 2005.

[2] C. Vidaurre and B. Blankertz, "Towards a Cure for BCI Illiteracy," *Brain Topography*, 2009 (ahead of print).

[3] B. Blankertz, G. Dornhege, C. Shafer, R. Krepki, J. Kohlmorgen, K. Muller, V. Kunzmann, F. Losch, and G. Curio, "Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis," *IEEE Transactions on Neural Systems and Rehabilitative Engineering,* vol. 11, no. 2, pp. 100-104, 2003.

[4] Compiled from http://www.bbci.de/competition/

[5] H. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularized common spatial patterns with generic learning for EEG signal classification," *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society*, 2009.

[6] Y.P.A. Yong, N.J. Hurley, and G.C.M. Silvestre, "Single-trial EEG classification for brain-computer interface using wavelet decomposition," presented at *EUSPCO conference,* Sept. 2005.

[7] S. Li and C. Shao, "Classification of Single Trial EEG Based on Cloud Model for Brain-Computer Interfaces," *Lecture Notes in Computer Science*, vol. 4689, pp. 335-343, 2007.

[8] F. Lee, R. Scherer, R. Leeb, C. Neuper, H. Bischof, and G. Pfurtscheller, "A comparative analysis of multi-class EEG classification for brain computer interface," in proceedings of *10th Computer Vision Winter Workshop*, 2005, 195-204.

[9] See μ rhythm tutorials at www.bci2000.org/wiki/

[10] J. R. Wolpaw, D. J. McFarland, G. W. Neat and C. A. Forneris, "An EEG-based brain-computer interface for cursor control," *Electroencephalogr. Clin. Neurophysiol.,* vol. 78, pp. 252-259, Mar. 1991.

[11] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm