Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms

Dornhege et al.
Berlin BCI Group

Goals of this Work

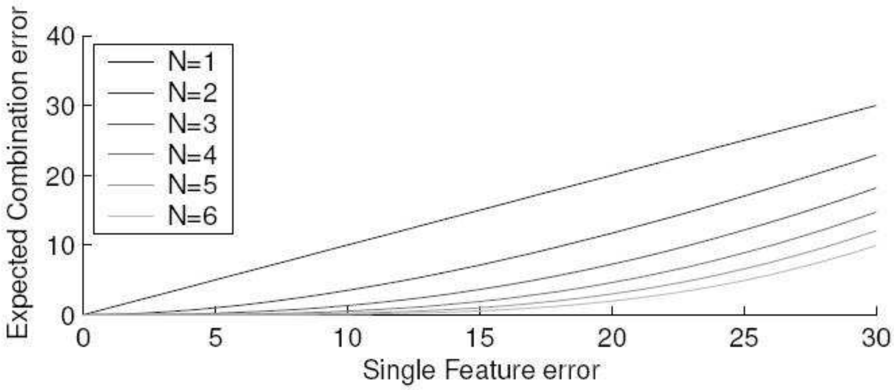
- ITRs need to be increased in order to make real uses of BCI possible.
- Explore the use of multiple features
  - Derive features from ERD, MRP, etc.
- Explore the use of multiple classes
  - More bits transferred per correct prediction.
  - Classifier accuracy could drop.

Obtaining Multiple Features

- ERD: Shows up as amplitude decrease of the α and β rhythms.
- MRP: Shows up as a negative DC shift, with a steep negative slope just prior to movement.
- Having multiple features can also provide better robustness against artifacts from outside CNS.

Why Having Multiple Features Helps?

- Assume Gaussian class-conditional densities for each feature X_j with means $(+\mu_j , -\mu_j)$ and variance σ_j^2.
- Error probability when using LDA classifier and feature X_j is $g(\mu_j / \sigma_j)$.
- Define $X = \Sigma_j (X_j / \sigma_j)$.
- Error probability when using LDA classifier and feature X can be shown to be smaller.



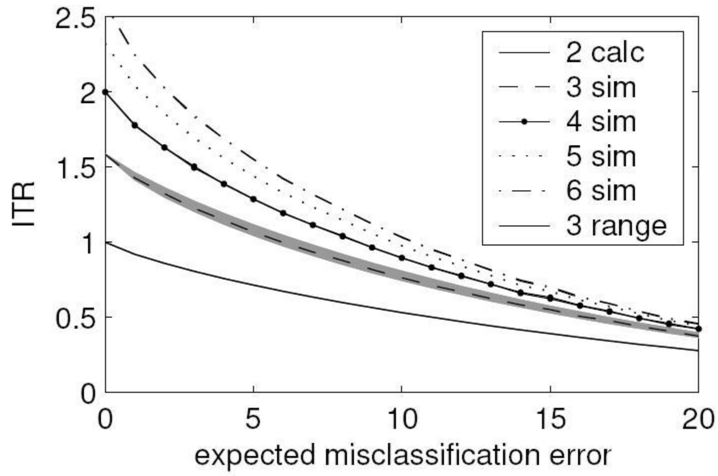
Why Having Multiple Classes Helps?

- Assuming Gaussian class-conditional densities, and a pairwise classification error *err* :

$$\frac{\exp(-\Phi(err)^2/6)}{6} \leq P(f^*(X) \neq Y) - err \leq \frac{\exp(-\Phi(err)^2/8)}{6}.$$

- If there are N classes, the classifier accuracy is p, X is the actual class, and Y is predicted class:

ITR = I(X,Y) = log N + p log p + (1-p)log (1-p)/(N-1)



Data Collection

- Use 64 or 128 EEG channels, band-pass filtered between 0.05 and 200 Hz.
- EMG and EOG measurements were done, but contaminated trials were not rejected.
- Had up to 6 different mental tasks – 3 motor imagery, and 1 each of visual, auditory and tactile imagery.
- 160 – 200 trials done for each class.

Classification Methodology

- Use LDA classifiers in their experiments.
- Do LDA regularization to tackle the “curse of dimensionality”.
- Do 10-fold cross-validation 10 times.

Three Kinds of Features

- DC Potential Shifts (SUB):
  - Correspond to the negative DC shifts seen in MRPs.
- Autoregressive Models (AR):
  - Each point of a time series is represented as a fixed linear combination of the preceding p points.
  - The AR coefficients reflect oscillatory properties.
- Common Spatial Patterns (CSP):
  - Will talk about them later.

Combining Features

- CONCAT: The features are simply concatenated, and fed to the classifier.
- PROB: If $g_{j,y}(x_j)$ denotes the class-conditional density for feature X_j, assuming independence of the features, the class-conditional for the combined features will be $g_y(x) = \Pi_j \, g_{j,y}(x_j)$.
- META: The meta level classifier is fed the results of classifiers working on individual features.

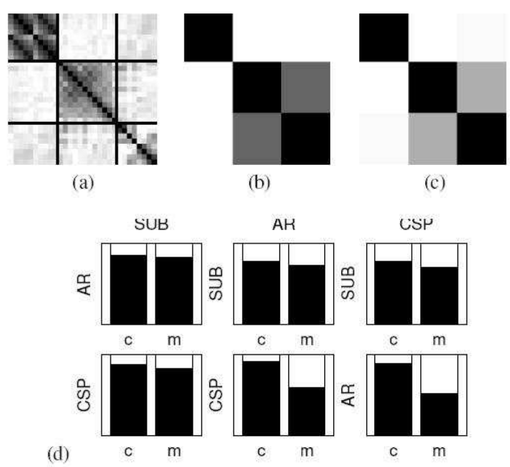


Fig. 4. In subplot (a) the absolute values of the correlation matrix for subject *aa* for (selected dimensions of) the feature-concatenated vectors are plotted as an image, where white points are values close to zero. The block order from top to bottom and left to right is SUB, AR, CSP. The two 3×3 matrices in (b) and (c) visualize the correlation matrix between the feature vectors (b) with respect to the real-valued leave-one-out output of each trial and (c) with respect to the leave-one-out error vector for each trial. In both cases the

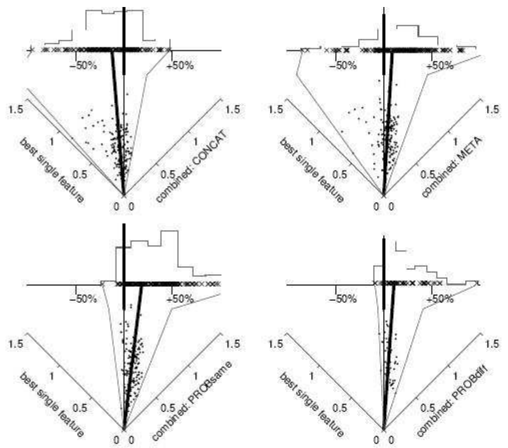


Fig. 5. The visualized scatter plots show the ITR on the best single feature vector based classifier against the presented combination methods for all 2 and multi-class combinations of all experiments except the ones described above. Above each scatter plot a histogram of the increase in percent in ITR compared to the best single feature vector is plotted. For points right of the vertical line through 0 in each scatter plot the combination algorithm outperforms the best single feature vector. The fat line shows the regression line of the points through the zero point calculated by minimizing the squared

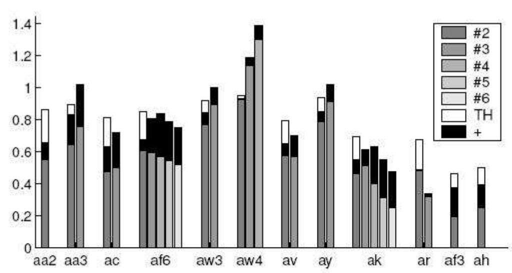


Fig. 7. The bar plot visualizes the highest ITRs for all algorithms without combination (all colors except black) presented here for all subjects for different numbers of classes from 2 (dark gray, #2) to 6 (light gray, #6). As a prolongation of each color bar we show the performance gain achieved with a combination method in black (+) and for the two class subsets the gain in white (TH) which theoretically can be achieved by formula (1) if feature vectors are perfectly independent. The number behind the subject code specifies the number of classes used for the specific experiment for subjects who took part in more than one experiment.